DOCUMENT RESUME

ED 348 402 TM 018 899

AUTHOR Bateson, David J., Ed.
TITLE Classroom Testing in Canada. Proceedings of the Canadian Conference on Classroom Testing (2nd, Vancouver, British Columbia, Canada, June 1-2, 1990).
INSTITUTION British Columbia Univ., Vancouver. Centre for Applied Studies in Evaluation.
REPORT NO ISBN-0-88865-195-3
PUB DATE Mar 92
NOTE 119p.
PUB TYPE Collected Works - Conference Proceedings (021)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Classroom Techniques; *Educational Assessment; Elementary Secondary Education; Foreign Countries; Group Discussion; Program Implementation; *Student Evaluation; Teaching Skills; Testing Programs; *Test Use
IDENTIFIERS Authentic Assessment; *Canada; Curriculum Based Assessment; Performance Based Evaluation; Schema Theory

ABSTRACT

This document provides the following 13 conference papers: (1) "Introduction: Where Do We Go From Here?" (D. J. Bateson); (2) "The Context of Classroom Procedures in Evaluating Students" (R. J. Wilson); (3) "Student Evaluation in the Ungraded Primary School: The SCRP (Systematic Cumulative Record of Performance) Principle" (L. McLean); (4) "The Assessment of Group Discussions and Complex Problem Solving: Potential Contributions of Schema Theory" (P. Nagy); (5) "Construction of Curriculum Relevant Tests by Teachers and Experts" (B. S. Randhawa); (6) "Considerations for the Implementation of an Ungraded Primary Program" (K. A. MacRury); (7) "What Should a Classroom Testing Program Look Like? The Functional Factors of an Assessment Program in Primary Classrooms" (J. O. Anderson and D. G. Bachor); (8) "Classroom Assessment: What Research Do Practitioners Need?" (I. McIntyre); (9) "Emerging Needs of the Practitioner in B.C. (British Columbia) Classrooms" (A. R. Taylor); (10) "Grounded Authentic Assessment and Teacher Education" (T. O. Maguire); (11) "What Skills Do Teachers Need in Educational Testing?" (R. K. Hambleton); (12) "Making Assessment Training Relevant for Teachers" (R. J. Stiggins); and (13) "A Call for Measurement Standards in Canada" (W. T. Rogers). (RLC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED348402

Classroom Testing in Canada

Proceedings of the Second Canadian Conference on Classroom Testing · June 1 and 2, 1990

Edited by
David J. Bateson

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MICHAEL MARSHALL

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

The Centre for Applied Studies in Evaluation (CASE)

The University of British Columbia

TM018899

2 BEST COPY AVAILABLE

Classroom Testing in Canada
edited by
David J. Bateson

Proceedings of the Second Canadian Conference on Classroom Testing held at the University of British Columbia, Vancouver, B.C. June 1 and 2, 1990.

This conference was supported by an occasional conference grant from the Social Sciences and Humanities Research Council of Canada (SSHRC), and the Department of Mathematics and Science Education (MSED) and the Centre for Applied Studies in Education (CASE) at the University of British Columbia. Formatting and production of this document were done under the direction of Michael Howell-Jones, Audio Visual Services, Faculty of Education, University of British Columbia.

© The Centre for Applied Studies in Evaluation (CASE), U.B.C.

March, 1992.
ISBN 0-88865-195-3

The Centre for Applied Studies in Evaluation (CASE) is a research centre and project management support centre within the Faculty of Education, University of British Columbia. CASE is an evaluation research centre which provides resource planning, scheduling, budget administration, record keeping, and overall coordination of projects whose focus is on evaluation.

CASE also provides evaluation research opportunities and financial support to graduate students and visiting scholars, acts as a 'broker' for individuals or agencies seeking evaluation project opportunities, and conducts seminars/ workshops with an evaluation focus.

BEST COPY AVAILABLE

Classroom Testing in Canada

Table of Contents

Introduction: Where Do We Go From Here? 1
David J. Bateson

The Context of Classroom Procedures in Evaluating Students 3
Robert J. Wilson

Student Evaluation in the Ungraded Primary School:
The SCRP Principle 11
Les McLean

The Assessment of Group Discussions and Complex
Problem Solving: Potential Contributions of Schema Theory 20
Philip Nagy

Construction of Curriculum Relevant Tests
by Teachers and Experts 38
Bikkar S. Randhawa

Considerations for the Implementation of
an Ungraded Primary Program 51
Katherine A. MacRury

What Should a Classroom Testing Program Look Like?
The Functional Factors of an Assessment Program in
Primary Classrooms 59
John O. Anderson and Dan G. Bachor

Classroom Assessment:
What Research Do Practitioners Need? 69
Iris McIntyre

Emerging Needs of The Practitioner in B.C. Classrooms 75
Alan R. Taylor

Grounded Authentic Assessment and Teacher Education 82
Thomas O. Maguire

What Skills Do Teachers Need in Educational Testing? 91
Ronald K. Hambleton

Making Assessment Training Relevant for Teachers 97
Richard J. Stiggins

A Call for Measurement Standards in Canada 104
W. Todd Rogers

Introduction: Where Do We Go From Here?

David J. Bateson

In June of 1990, the Second Canadian Conference on Classroom Testing was held at the University of British Columbia, the first having been held at the University of Victoria in 1989. Many of the papers from that first conference were published in a special issue of the *Alberta Journal of Educational Research*. This present document is a compilation of most of the papers which were presented at the second conference.

The conference was held at a time when educational measurement, and classroom testing in particular, was undergoing a radical change in both theory and practice. The introduction of the concept of "authentic" measurement in *Beyond Standardized Testing* (Archbald & Newmann, 1988) has provided the stimulus for an enormous shift in the conduct of educational measurement as it is practiced in classrooms. This shift has been demonstrated in everything from the content of the annual meeting program of the National Council for Measurement in Education to the documents surrounding the *Year 2000* initiatives in British Columbia.

In order to facilitate discussion, participants in this conference were sent copies of *Year 2000: A Curriculum and Assessment Framework for the Future* (Ministry of Education, 1989a) and *The Primary Program* (Ministry of Education, 1989b), the draft proposal for the announced changes to the primary program in British Columbia, as a stimulus for conference papers. These documents include recommendations and even demands that teachers work toward "authentic" assessment procedures (1989a, p.17). They include quotations such as "assessment and evaluation techniques must mirror the actual learning experiences of the child..." (1989b, p.13), "for most assessment purposes in school, traditional forms of standardized tests are not very useful, and they are inappropriate in the Primary Program..." (1989a, p.17), "there is a shift from examination to demonstration..." (1989b, p.13), "comparison with other students, or the assignment and use of letter-grade or pseudo letter-grade symbols is also inappropriate for primary children...", (1989b, p.13) and "although checklists are useful for teachers as a way of organizing information, they are not appropriate as reporting devices..." (1989b, p.166).

The statements from the *Year 2000* initiatives are representative of much of the discussion regarding classroom-based educational measurement in Canada at the present time. Many of these evaluation issues which are proliferating the present educational scene attack the roots of traditional measurement theory and practice. For at least the last two decades, most measurement specialists have concentrated their efforts on large-scale, standardized, selection-type assessment techniques; large-scale, "high-stakes" testing programs; and methods that

employ the massive number-crunching capabilities that the rapidly expanding computer technologies have made available. This situation is understandable since these are the areas where funding and other resources have been available from technology-based companies who are interested in marketing products and from politicians and governments who are obsessed with problems of accountability and public perception of declining standards. However, as educators, particularly classroom teachers, have become more and more disillusioned and disgruntled with what they perceive to be a warped picture of students and education in general which has resulted from the exclusive use of the type of information provided by these large-scale, standardized, selection-type tests, the need for a radical shift in thinking and practice has emerged.

Measurement specialists in Canada, rather than opposing many of the ideas of the "authentic" movement that attack past practice, have rapidly confirmed and endorsed the basic philosophy of these ideas. However, these same specialists are working to ensure that new practices utilize what is already an extensive theory and knowledge base of educational measurement, rather than ignoring all that has gone before and starting anew. It is essential that a destructive revolution in educational measurement be avoided. What is necessary is a rational, but fairly rapid evolution. Within this context, the Second Canadian Conference on Classroom Testing was convened, and it is hoped that this collection of papers can contribute to the present evolution in measurement theory and practice.

References:

Archbald, D.A. & Newmann, F.M. (1988). *Beyond Standardized testing: Assessing authentic academic achievement in the secondary School.* Reston, VA: National Association of Secondary School Principals.

Ministry of Education. (1989a). *Year 2000: A curriculum and assessment framework for the future.* Victoria: Author.

Ministry of Education. (1989b). *The Primary Program.* Victoria: Author.

Dr. David J. Bateson, the organizer and chairperson for the conference, is an Associate Professor in the Department of Mathematics and Science Education at the University of British Columbia and a Research Associate with both the Educational Measurement Research Group (EMRG) and the Centre for Applied Studies in Evaluation (CASE).

THE CONTEXT OFCLASS ROOM PROCEDURES IN EVALUATING STUDENTS

Robert J. Wilson
Queen's University

Much of the attention devoted to classroom-based assessment practice concentrates on the teacher: what the teacher does, what the teacher might need, and what the teacher should do.

Various suggestions have been made to teachers. In recent curriculum documents teachers are advised to use a wide range of assessment tools, concentrating on observation and process measures, and to store the records of individual accomplishments in student development files (B.C. Ministry of Education, 1990). Texts for courses in educational measurement generally feature adaptations of psychometrics and professional practice to classrooms (Brown, 1981; Cunningham, 1986). Other writers, convinced that the adaptive method has not adapted well enough, have used a comparative analysis of classroom practice and measurement standards to recommend where and how a new melding of theory and practice might profitably occur (Frisbie and Friedman, 1987).

Two assumptions about classroom-based assessments are usually made in these efforts. First, the classroom teacher is performing the evaluation activity *primarily* to inform judgements about student progress so that instruction and learning can proceed more effectively. Secondly, the teacher is virtually a free agent in determining which particular evaluation activities and purposes will be adopted in the classroom.

The purpose of the present paper is to investigate these two assumptions by referring to work we have done and are presently doing that is aimed at exploring the broader environment in which student assessment occurs. [1]

What do Teachers Want to Do?

That teachers wish to fulfill the first assumption about the value of evaluation has not been validated by much research. We are presently examining the attitudes of 101 Bachelor of Education students during their year of teacher preparation. As part of their participation in this study, these prospective teachers were individually interviewed halfway through this year with an instrument designed to explore the reasons for the attitudes they hold toward evaluating students. (They also completed a Likert-type attitude scale on three separate occasions during the year.)

[1] This work, supported by the Social Sciences and Mathematics Research Council, includes among its members, Ruth Rees, Marilyn Cannock, Lyn Shulha, Alison Taylor, besides the author.

One of the items in the interview asks them to rate the importance of various purposes for administering evaluation instruments to children. The student teachers' results (transformed into ranks from most important to least important) are provided in Table 1 displayed by level of intended practice. Students preparing to teach from Kindergarten to Grade 6 are labelled PJ (Primary-Junior). Those preparing to teach from Grades 4 through 10 are labelled JI (Junior-Intermediate), and those preparing to teach from Grade 9 through Ontario Academic Course (formerly Grade 13) are labelled IS (Intermediate Senior).

Table 1

A Comparison Among Student Teachers Concerning Importance of Various Reasons for Evaluating Students.

(Data Given in Ranked Order) Reasons	PI's	JI's	IS's
To check students' progress against course objectives	2	1	1(tie)
To compare students' achievement to others	10	10	10
To generate marks for reporting purposes	9	9	9
To ensure that students do assigned work	7	8	7
To prepare students for this kind of evaluation in the future	8	7	8
To have students practice or apply what has been learned	6	4	5
To diagnose students' weaknesses with the material	3(tie)	3	3
To enable students to monitor their own progress	5	2	1(tie)
To help me decide what to teach next	1	6	6
To allow me to see how well I taught the material	3(tie)	5	4

The data indicate that these student teachers tend to agree with each other on the relative importance of various purposes of evaluation no matter at which level within the system they intend to work. Key differences seem to exist between the PJ students and the other two groups on the use of evaluation results "to help me decide what to teach next" and "to enable students to monitor their own progress," but on most of the suggested purposes, little disagreement emerges.

The highest ranked purposes for all three groups can be interpreted as a general tendency to see evaluation activity as informing the teaching-learning process. Purposes with this overall function in mind (objectives-reference, applying learning, diagnosis, self-monitoring) are more highly weighted than those whose functions is more administrative in nature (generate marks, prepare for future exams, norm-reference).

In this respect, these student teachers agree with their colleagues on curriculum-writing teams, evaluation text authors, and instructors in measurement courses concerning the major purpose of evaluation in classrooms: to inform the teaching-learning process.

What Do Teachers Actually Do?

The group they do not agree with so well are their more experienced colleagues. Virtually the same item was used in a previous study (Wilson, 1990) in which 51 practising teachers in British Columbia and Ontario filled in the same scale (presented as a check-list of purposes) for each instrument they administered to their classes during a reporting period. The 24 teachers comprising the Ontario sub-sample were divided into two groups of 8 elementary (Kindergarten through Grade 6) and 16 secondary (Grade 9 through Grade 13) teachers. Table 2 shows the comparisons of the ranked frequency of purpose of the practising teachers with those of the relevant student teacher comparison groups.

Compared to all the student teachers, the practicing teachers rate "the generation of marks for reporting purposes" much higher. Indeed, this purpose dominates the entire exercise for the experienced teachers, it having been checked as a purpose on four out of every five instruments. Other disparities between the practicing teachers and the student teachers grow larger as the level of intended practice increases. The differences are most pronounced at the secondary level where the highest ranked purposes for evaluation given by the practicing teachers refer to administrative and external aims, all of which are the lowest ranked purposes for the prospective teachers.

Our study of these 101 teachers is continuing. We will interview as many of them as become employed in the province of Ontario twice more during the coming school year. In this way we hope to determine how they adapt their present views on assessment to the reality of their own classroom life and also try to determine why their views alter if they do.

Table 2

A Comparison of Practising Teachers (PT's) and Student Teachers (ST's) Concerning Importance of Various Reasons for Evaluating Students.

(Data Given in Ranked Order)

Reasons	PT's		ST's		
	ELEM	SEC	PI	JI	IS
To check students' progress against course objectives	4	2	2	1	1(tie)
To compare students' achievement to others	10	6	10	10	10
To generate marks for reporting purposes	3	1	9	9	9
To ensure that students do assigned work	9	4	7	8	7
To prepare students for this kind of evaluation in the future	8	5	8	7	8
To have students practice or apply what has been learned	2	3	6	4	5
To diagnose students' weaknesses with the material	1	8	3(tie)	3	3
To enable students to monitor their own progress	5	7	5	2	1(tie)
To help me decide what to teach next	7	10	1	6	6
To allow me to see how well I taught the material	6	9	3(tie)	5	4

What Accounts for the Differences?

Through other work we are doing (Rees, 1989; Wilson and Rees, 1990), we expect that their views will alter. We have hypothesized that the policies and procedures concerning student achievement devolved upon teachers from levels "above" them in the administrative hierarchy will force their evaluation activities into relatively narrow areas.

Table 3

Content of Policies and Procedures in Student Achievement by Level of Origin.

Content	Level				
	Ministry	District	School	Dept.	Classroom
Scholarships	X				
Consultation re Evaluation	X				
Appeals	X				
Reporting, Grading	X	X	X	X	X
Individual Assessment	X	X	X	X	X
Examinations	X	X	X	X	X
Promotion	X		X		
Attendance			X	X	X
Communication to Students			X	X	X
Weighting of Evaluation Types			X	X	X
Timing and Types of Evaluation				X	X

Table 3 shows the policy content and the level at which this content becomes part of the student assessment environment for teachers in two districts and several elementary and secondary schools in British Columbia. (Our analysis

of the Ontario environment is not yet completed.) There is a very heavy emphasis in the policy and procedural framework on examinations, reporting formats, promotion, and evaluation weights. Much of the content of these policies creates a standardization of approach (common examinations, set timelines for reporting, and mandated reporting symbols) that forces the assessment activities of teachers onto relatively restricted and narrow paths.

Our study of the policy area of student achievement concludes that there is a mismatch among (a) the few policy statements that provide philosophical direction to teachers; (b) the actual procedures that teachers and others (particularly beyond the primary grades) must implement; and (c) overall educational goal statements for public education in the province (Wilson and Rees, 1990).

For example, during the 1987/88 school year, the B.C. Ministry of Education had a policy which advocated the following goal: "It is essential that parents be kept informed of progress achieved toward expectations held in common by the teacher, student, parents, and community." In the actual implementation of reporting to parents, however, all the schools in our sample beyond the primary level provided relative judgements, usually in letter grade format, that defined expectations in percentage range equivalents. (In fact, such a definition is completely circular as neither letter grades nor percentages are defined in terms other than each other.)

The specificity and standardization of most of the procedures at the school level, and the linking of these procedures to specific calendar dates, ensures an attention to a school-year rhythm based on the reporting cycle. It might be assumed, then, that teachers could fulfil the reporting requirement with instruments that were easy to prepare and mark but which may or may not be related to the actual learning going on in the classroom, particularly if there were no other clear expectations for the evaluative process. What is more serious, however, might be the possibility that a low expectation level for the evaluation gradually comes to replace a more ambitious expectation for the actual learning.

Other investigators (Carter, 1984; Stiggins, Griswold, & Wikelund, 1989) have shown that the level of assessment typically employed by teachers at all levels is quite low in terms of cognitive complexity. In our data, it is clear that the cognitive demand of single-word completions and short-answer items (questions which appeared on 44% of the instruments we collected) is not likely to be high if for no other reason than that the format does not allow any higher level than recall of specific bits.

The same case for multiple-choice questions, and for other supply and performance items, is not so easily made. Here, for example, is a multiple-choice item taken from a Grade 12 examination:

1. What is the distance between the points (1, -8) and (-4, 4)?

   A. 5
   B. 7
   C. 13
   D. 17

It may be that the student faced with this item selects a formula from

memory, substitutes the relevant data from the stem, and selects the answer closest to his calculation. It may also be the case, however, that a student without that formula, but with a knowledge of the Pythagorean theorem, applies that knowledge to the same request and responds in exactly the same way as her fellow student. The conclusion to be made is that it is difficult to determine from questions phrased in this manner **anything** reliable about the operations actually used by students to respond to the item.

Two conclusions seem fair for the types of evaluation instruments we found in our earlier work. First, many items allow students to demonstrate only a low level of cognitive operations. Second, those that do allow for more complex operations may not provide unambiguous data about those operations.

While it would seem obvious that recommendations to develop items that do encourage better interpretation would be welcome (the work of Biggs and Collis (1982) is especially noteworthy here), such recommendations assume that the present model is not meeting teachers' needs. But if the teachers' real needs are to meet reporting demands for single label judgements of students' relative standing, and to accomplish that task with a minimum of time spent on it, then recommendations for more involved development, scoring, and interpretation of complex items will seem, at best, irritating and irrelevant. Perhaps the reason why the cognitive level exhibited in classroom assessment instruments does not alter much through the school cycle is that such growth is neither required nor expected by the evaluation policy of the school.

Conclusion

These results, admittedly fragmentary and preliminary, may nevertheless provide a cautionary note: Before those of us interested in the evaluation activities of teachers in classrooms proceed too far down a road toward implementation of newer approaches, we might first attend to the broader environment in which these activities occur.

It may be that teachers use the results of classroom assessments for reasons other than those posited by documents, experts, and naive practitioners. Unless the environment of classrooms can be altered so that certain administrative functions concerning reporting, attendance, and communication with outsiders are less overwhelmingly intrusive, it will be non-productive to work with teachers alone to change their present practice. The teachers, at least those we have worked with, are not free agents to make the types of changes outsiders deem desirable.

It was a major breakthrough to understand that the classroom itself has a life that shapes many of the activities that go on there, including evaluation of student learning. Now it seems that a classroom can be seen as a unit in a larger structure which also creates a community, a community with very intrusive expectations. Attending to that larger unit, and altering what it considers necessary and desirable, may well be a prerequisite to successful change in student assessment practices at the classroom level.

References

B.C. Ministry of Education, (1990). *A program for schooling in British Columbia: The intermediate years*. Victoria: Ministry of Education.

Biggs, J.B., and Collis, K.F. (1982). *Evaluating the quality of learning*. New York: Academic Press.

Brown, F.G. (1981). *Measuring classroom achievement*. New York: Holt, Rinehart and Winston.

Cunningham, G.K. (1980). ***Education and psychological measurement***. New York: Macmillan.

Frisbie, D.A., & Friedman, S.J. (1987). Test standards: Some implications for the measurement curriculum. *Educational Measurement: Issues and Practice*, 6(3), 17-23.

Rees, R. (1989). *The influence of the provincial educational system on educators' evaluation policies and practices*. Paper presented at Ontario Educational Research Council Annual Conference, Toronto.

Stiggins, R.J., Griswold, M.M., and Wikelund, K.R. (1989). Measuring thinking skills through classroom assessment. *Journal of Educational Measurement*, 26(3), 233-246.

Wilson, R.J. and Rees, R. (1990). *Characteristics of policy and procedures governing student evaluation*. Paper presented at the Annual Meeting of the Canadian Society for the Study of Education, Victoria.

Student Evaluation in the Ungraded Primary School: The SCRP Principle

Les McLean
Ontario Institute for Studies in Education

Ungraded schools follow naturally from other developments that have been shaping primary schools in recent years. Continuous progress has always been a fact of school life, but lately it has been given official status and sanction. The idea of curriculum keyed to school years is inconsistent with the approach to the teaching of reading, writing, listening and speaking known as *whole language pedagogy*, and all sit within a still evolving concept called *activity based learning*. Implement these with a policy that students with handicaps of various kinds will be taught in regular classrooms as much as possible (mainstreaming), and schooling organized by school year must finally be abandoned.

With the departure of grade levels, many traditional ways of defining and evaluating student achievement go as well. The most obvious is the grade equivalent score. Grade equivalents in reading have been criticized by language theorists and curriculum specialists for a decade, so their demise is to be welcomed. Their shaky statistical properties and poor substantive rationale make them suspect in every subject. A major problem is that their existence and use for so many years has created the belief that achievement in the primary school can be captured by numbers accurate to two or three significant figures. "Jane's reading level is 5.4, and she is only in Grade Three!" A similar belief in precise numbers has led to the creation of other test scores using the so-called, "item response theory". When the Americans came up with the name, "Rit" for their scale unit, far western Canada was ready with its "Brit". Alas, school achievement is much more complex than that — and considerably less precise.

Grade equivalents will persist (with Brits and other fictions) as tools to bring order out of chaos until we can offer teachers and officials something good and practical to replace them. The purpose of this paper is to suggest a principle and some implications for good and practical replacements.

The SCRP Principle

Measurement specialists have long advocated testing as the main tool for student evaluation, and teachers do construct and administer many tests. In the primary school, however, tests have never been teachers' evaluation method of choice. Observation and informal evaluation of student work are listed as the dominant methods (Wahlstrom, 1977, Fair et al., 1980). At their best, these methods provide authentic assessment of achievement. Unfortunately, they are not always at their best, and neither classical nor modern test theory has had anything to offer in suggesting improvements. The challenges of informal evaluation are not best described in terms of reliability and validity; better concepts are fairness and thoroughness. It is very difficult to observe and

evaluate informally in an active primary school classroom and still ensure that students are treated evenhandedly and that every student's work is evaluated.

The complexity and difficulty of primary classrooms rule out simple, uniform procedures of any kind, including evaluation methods. No one approach will work in all classrooms. In such a situation, we usually turn to an organizing principle that can guide decision making while preserving flexibility. Teachers have to make thousands of choices per day under only loosely controlled conditions. *Whole language* teaching is based on such a principle, with some detailed procedures worked out in advance and many decisions made on the spot. The guiding principle is that language is not divisible, that is, reading, writing, listening and speaking are interconnected and dependent in some known and some still mysterious ways. Such a principle has profound implications for pedagogy, requiring radically different teaching and testing methods than were the consensus choice only a decade ago (McLean, 1988).

Whole language teaching and testing emphasizes the communication of meaning in context. Assessment tasks must therefore present a context, making them longer and more complex. Students bring different backgrounds to the tasks, and this natural variation reduces traditional reliability indices (such as internal consistency). In second language testing, it seems that the better the test (the more it reflects communicative language theory) the *lower* will be the value of Cronbach's Alpha (Swain, 1990). Assessment that faithfully reflects student performance in meaningful contexts is referred to as "authentic", a quality concept closely related to validity but better suited to achievement testing (Archbald & Newmann, 1988). Most test development effort in North America and Europe is now going into tasks requiring students to demonstrate abilities and skills directly, in other words, by *performance*. Directly interpretable evidence is proving to be more useful and attractive to educators than indirect and abstract test scores.

Activity based learning introduces another aspect that should be mentioned before considering assessment principles. In order for activity based learning to succeed, the teacher has to establish an atmosphere of respect and trust in the classroom. Students have to agree to listen to each other, to praise and to criticize with respect. One particularly successful teacher spends much of the first two weeks of the school year building such an atmosphere by setting out some rules, getting students to propose others and by starting activities that illustrate the need for such a spirit (McLean, Aitken, Van Duzer and Peterson, 1990). In the beginning, the teacher will assign students to work together on a common task in small groups. These are not the familiar redbirds, bluebirds and other reading groups, which are a form of ability grouping. Each group should reflect the ability range in the classroom as closely as possible. Soon, students are working in different groupings as tasks change, or working alone on an assignment. Classroom atmosphere is a topic in itself, but here we have to move on with this brief mention.

The assessment principle suggested to go with these changes is that evaluation should be based on a *Systematic Cumulative Record of Performance*,

SCRP. Each of the words has been carefully chosen in light of the more abstract and general principles of fairness and thoroughness, so each will be discussed in turn. The discussion is smoother if we take them in reverse order. The setting for the discussion is assumed to be primary schools — the first three or four years of schooling after age 6.

Record of Performance

Evaluation can and should be based on actual samples of students' work — on their performance of meaningful tasks. Teachers in the early years do much of this now, with precious little guidance how it might be structured and fairly done. Six-year-olds draw pictures and dictate the meaning to their teacher or other adult, for example, and the adult writes the "story" on the picture (or on the back). Words should always be associated with a story, and stories told are valued as much as stories read. One theorist argues that stories told are especially good *bonnes a penser*, food for the mind, and that stories are the building blocks of a complete elementary school curriculum. For evaluation, "one might particularly focus on ... something written, dramatized, drawn, that gives evidence of the effective effect of the unit while drawing on supporting knowledge, skills, and so on" (Egan, 1988, p. 247).

Performance on meaningful tasks should be captured in portfolios, records of achievement retained and managed by each student. Note that it may be a record of performance that is retained by the student, rather than the performance itself. When something is "dramatized", for example, the performance may not be retained at all. Video cameras are becoming more and more common in elementary schools, so even the dramatizations can be captured, but the video cassette need not be kept in the portfolio. The existence of the cassette, and perhaps its location, is recorded in the portfolio, because the portfolio is the basis for both formative and summative evaluation. Students will have two classes of portfolios — active and cumulative. Current projects, works in progress, are kept in the active portfolio, and samples to be retained in the record are transferred to the cumulative portfolio when ready to move on to another topic. When working in a group, each student makes a copy of her or his performance for inclusion in their record. Performance is often a group effort, but individual accountability is associated with better learning (Slavin, 1987).

Such great reliance on portfolios dictates that there be some backup. What if a student loses her or his cumulative portfolio near the end of a term or school year? As discussed in detail below, evaluation of portfolios is a periodic joint effort of teacher and student. At evaluation time, the teacher should make and retain summary notes, the main purpose of which is backup in case the actual record is lost. The teacher's notes would be a poor approximation, but with the student's help, they could be used to reconstruct a reasonable summary of achievement up to the time of latest review. A teacher's note might look something like this, the student having seen (or heard) and agreed to it.

RoA reviewed with ___ on __ (date) __. The record was up-to-date in math and science. Quality was uneven, with a few excellent and a majority good

completed tasks.[1] One or two were poor, and plans to improve them by the next review. Reading list was well filled, but writing samples were almost lacking. Mark at present was agreed to be marginal to good.

Most of the research and development on records of achievement has been done in the United Kingdom, starting more than 10 years ago in Scotland. Records became educational policy a few years ago, but the national curriculum and national testing have thrown doubt on their future. A comprehensive evaluation was carried out in local authorities across England and Wales, and the report of that evaluation gives both history and practical information about strengths and weaknesses (Broadfoot, 1988). All may not be lost, however, because evidence just appeared that records have wider application than first thought. The department of education at the University of Cambridge has introduced records of achievement for about 60 of their Postgraduate Certificate in Education students as part of their preservice training. The same benefits are seen for these graduates preparing to enter teaching as for elementary and secondary school students.

We hope that these profiles will help students to set goals for themselves, to see what they have learned and to identify the areas in which they can develop further. We also expect that by compiling a profile, a student will get an idea of what it is like for children to complete records of achievements. It will also give students a flavour of what teacher appraisal may entail when it eventually arrives (Beardon & Reiss, 1990).

In the U.S., 50-75 colleges and universities are using them to evaluate the teaching performance of professors (Watkins, 1990).

What we see emerging is that evaluation can finally become an integral part of teaching and learning, and that evaluation, teaching and learning can all be usefully conceived within theoretical frameworks based on language as meaning. Theoretical frameworks not only tell us how to organize the teaching but also how to structure and score the tasks. Teaching may yet get beyond craft and become a profession (McLean, 1985).

Cumulative Records

Research in cognitive science has illuminated many aspects of meaningful learning, and one aspect that stands out is the emphasis on the cumulative nature of learning. New material must be integrated with what we already know if it is to be remembered and applied. Indeed, since the need to relate to prior learning is so strong, it is a paradox that learning gets started at all. The whole concept of *meaning* is that we make sense of novel information by relating it to what we already know. How does learning get started? It is especially clear that language competence is an outcome of a slow cumulative process, much of which is still mysterious.

[1] This work, supported by the Social Sciences and Humanities Research Council, includes among its members Ruth Rees, Marilyn Connock, Lyn Shulha, and Alison Taylor besides the author.

An unconnected and chaotic pile of performance records cannot, therefore, provide an authentic achievement measure, even if the tasks individually are authentic. Much of the learning that has taken place will be missed unless the students identify and explicitly record the links among their performance samples. This may be a case of reactive measurement, in that *the conscious search for links may identify connections which would otherwise go unnoticed.* If so, then so much the better, since learning is the objective, not some sort of ideal measurement. Cognitive scientists have also emphasized the crucial role reflection can play in learning. Stopping and "looking back", that is, thinking back, is an important step in consolidation of learning. Building their cumulative portfolio can give primary school pupils early practical practice in reflection and taxonomy-building.

Until students get used to keeping portfolios, teachers will have to spend some time explaining and illustrating their use. The mechanics of the active and cumulative folders are simple enough, but considerable discussion will be required about the cumulative folder. The teacher explains that when work on a topic or project is finished (or stopped, at least), a few samples of the student's best work should be saved. The work saved should show what has been accomplished, what solutions were obtained *and what it means.* Primary teachers will need a repertoire of examples to get the process started.

Students illustrate the meaning of a performance sample during a review by showing how the sample is linked to work already in the portfolio. Obviously, this evolves slowly for six-year-olds. Sometimes when new work is started, links will be few and difficult to show, and this work will be noted as needing more explanation later. The word "novel" can be introduced after a year or two, in the sense of "new; of a new kind or nature; hitherto unknown". Students learn that what is novel to one person may be well-known to others. It follows that every new sample added to a portfolio is an occasion for reflection.

The Importance of Being Systematic

Systems are required to deal with the complexity of primary classrooms, and teachers are good at figuring out systems. They have their record books, notebooks, folders and the like. What few teachers have is a schedule for evaluation and a system of recording that can ensure even and thorough coverage of their class(es). Teachers already have to assume more of a managerial role in activity based learning, and they quickly learn that when managers have too much to do, they delegate. In this case, teachers delegate the first level of record keeping to the students by means of portfolios. Two types of systems are needed for evaluation, (a) a way of describing content expectations, and (b) a schedule. We assume for the moment that students are already familiar with the purpose and operation of portfolios, and that the teacher has been able to establish an atmosphere of respect and cooperation.

Content expectations. The essence of ungraded schools is that students start at different stages and proceed at different rates; there are no fixed expectations by school year. The teacher must have expectations, of course, and ideally these should be tailored to each student. Operationally, this is a nightmare, and it

becomes worse as students diverge in their performance over several years. Here is where portfolios and the SCRP principle can be applied to good effect. Expectations are set in terms of cumulative growth.

Teachers communicate expectations to pupils and parents via general lists of the types of performance samples they wish to see in the portfolios. Examples are stories told, written and read (because some students are reading when they come to school), number exercises, science experiences and the like. The tailoring is done between teacher and student at review times, when specific targets are set for the time of the next review. Systematic, cumulative progress is the goal, and students readily understand it in principle. Understanding it in practice comes with time and with the opportunity to observe numerous examples. The challenge (nothing new) is for the teacher to set reasonable standards.

Standards are set in terms of starting points and benchmarks. They are communicated by means of narrative descriptions and sample portfolios. Teachers must be able to consult a collection of portfolios in the school or district office. Collections will include six and nine month portfolios (or other to match marking periods) for starting points such as these:

- Students who could read when they came to school, had English (or French, that is, the language of instruction of the school) as their first language and encountered no major obstacles adjusting to school. Included would be examples of slow, average and rapid progress. These are top benchmarks. Call them T1, T2 and T3.

- Students who did not read when they came to school but who were otherwise like the first group. Some will equal the top benchmarks in quality of performance at the end of the period. Call these M1, M2 and M3, the middle bechmarks.

- Students whose first language is not English or who have other problems getting started in school. Call these O1, O2 and O3, the "other" benchmarks.

The Board of Education for the City of Toronto has just published a comprehensive series of benchmarks for language and mathematics performance tasks after three and six years of schooling. The range of performance is huge, as you would expect, and show what happens even in "graded" schools. If they had portfolios, all three of the groups would be well represented. They illustrate how performance tasks can be used to suggest reasonable targets for students.

Review schedule. It is easy to work out that if a portfolio review takes 10 minutes, and there are 24 students in the class, four hours will be spent on each round of reviews. A teacher in primary can count on at most five hours per day of time suitable for reviewing, only 15 hours in three weeks. A quarter of class time will be spent reviewing portfolios. If there are 30 students, five hours will be required, accounting for one-third of class time. This can only be considered seriously if the time is also seen as prime instructional time, and that is precisely the claim that is made for it. The point here is that teachers have to plan and schedule reviewing at least carefully as they do any other activity.

Overview of SCRP

The SCRP principle sits in the mainstream of current developments in student evaluation - performance testing. The evidence is shifting decisively away from test scores to directly interpretable evidence in the form of performance on meaningful tasks. This is not a huge shift for primary school teachers, of course, because they have always depended on observations of performance in their evaluations. What is new is the recognition that evidence from performance testing can be organized and recorded to serve all the purpose that test scores served, but better. That said, it has to be noted that a traditional test can also be meaningful and that answers to the questions are a form of performance. Test papers that have been marked, perhaps annotated, are candidates for inclusion in records. Word recognition, spelling and arithmetic tests are not good candidates, however. They do not qualify as authentic measurement, and the time required to complete them is better spent in other ways.

Student evaluation serves multiple purposes. Its primary purpose is to supply feedback to students, teacher and parents, but increasingly it has to provide accountability to the wider community who pay the large cost of education. In order to meet all these demands, there must be a careful record of student performance and a way to interpret that record. The popularity of test scores can be explained in large part by the ease with which they could be recorded and stored and by the range of interpretation schemes devised for them by the testing profession. Only gradually did it become clear that the convenience and surface credibility of test scores were brought at the price of authenticity. Records of performance in the form of portfolios can provide authentic documentation of achievement and also provide summaries for purposes of accountability.

The key to pedagogical success is to ensure that the portfolios are cumulative, in the intellectual as well as the physical sense. This means that students and teachers work together to give meaning to the cumulative portfolio, asking where each new entry fits with the others and gradually constructing a content map of the record. Constructing the content map brings into play the powerful learning tool of reflection and ensures that the evaluation task is also a knowledge production task. The process can and should begin in the primary school, so that it becomes second nature to all students. Some of the most important teaching and learning will happen as students review their portfolios with the teacher and with other students.

Rich cumulative portfolios cannot emerge in large classes unless attended to systematically by teacher and students. The teacher has to establish an atmosphere of trust and sharing in which students work independently and in small groups without constant supervision. This means that the teacher is a manager, setting tasks and delegating most of the record keeping to the students. The teacher keeps a brief outline record from the review occasions, as a convenient reminder how far the process is with each student and as a backup in case a student's portfolio is lost. Only by systematic reviewing and record keeping can the teacher ensure that student evaluation is fair and thorough.

[1] A five-category scale is used for individual tasks, for example, Superior, Excellent, Good, Marginal, Unsatisfactory. Such outcomes can obviously be given numerical values and then weighted and summed or averaged if a quantitative summary is required for accountability purposes. Such a summary would never be very useful as feedback to students or in reporting to parents. Normative interpretations should be done in terms of number and kind of tasks completed in comparison to the rest of the class or to school expectations.

References

Archbald, Doug A. & Newmann, Fred M. (1988). *Beyond Standardized Testing: Assessing authentic academic achievement in the secondary school.* Washington: National Association of Secondary School Principals.

Beardon, Toni & Reiss, Michael (1990, April 20). A taste of things to come. *Times Educational Supplement* (London), p. 21.

Broadfoot, P., James, M., McMeeking, S., Nuttall, D. and Stierer, B. (1988). *Records of Achievement: Report of the National Evaluation of Pilot Schemes.* London: Her Majesty's Stationery Office.

Egan, Kieran (1988). *Primary Understanding: Education in early childhood.* New York: Routledge.

Fair, J.W., Biemiller, A.J., Grapko, M.F., Hunt, G.W., Martin, R.A., Mock, K.R., Pike, R., Sheehan, A.T., Volpe, R.J. & Wood, J.D. (1980). *Teacher Interaction and Observation Practices in the Evaluation of Student Achievement.* Toronto: The Minister of Education, Ontario.

McLean, Les (1985). *The Craft of Student Evaluation in Canada.* Toronto, Ontario: Canadian Education Association.

McLean, Leslie, D. (1988). Achievement measures made relevant to pedagogy. *The McGill Journal of Education, 23*(3), 243-252.

McLean, L., Aitken, J., Van Duzer, J. & Peterson, S. (1990, June 3). The role of verbalization in classroom learning. Paper presented at the Victoria, B.C. meeting of the Canadian Society for the Study of Education.

Slavin, R. (1987). Cooperative learning and the cooperative school. *Educational Leadership, 45*(3), 7-13.

Swain, Merrill (1990, March). Second language testing and second language acquisition: Is there a conflict with traditional psychometrics? Paper presented at the Round TAble on Language and Linguistics 1990—Linguistics, Language Teaching, and Language Acquisition: The interdependence of theory, practice and research meeting of the Georgetown University.

Wahlstrom, M.W. (1977). *Assessment of Student Achievement: A survey of the assessment of student achievement in Ontario.* Toronto: The Minister of Education, Ontario.

Watkins, Beverly T. (1990, May 16). New technique tested to evaluate college teaching. *Chronicle of Higher Education*, pp. A15-16.

The Assessment of Group Discussions and Complex Problem Solving: Potential Contributions of Schema Theory

Philip Nagy
The Ontario Institute for Studies in Education

Introduction

The purposes of this report centre on the issue of how to assess some of the more complex outcomes of education. The theoretical framework is based in the literature on the solving of ill-structured problems, in particular the application of schema theory to framing and comparing different solutions to such problems. I consider two examples: one, individual solutions of school principals to an administration problem presented as a case study (Nagy, 1990a); and two, discussions among groups of elementary school children concerning a family conflict situation, also presented as a case study (Nagy, 1990b). The concerns addressed are generalizable to the assessment of many complex educational outcomes, including expository essays, group discussions, and any problem solving situation where the notion of a simple marking scheme with one correct solution is inappropriate.

The shape of the paper is as follows. First, using the administrative problem example, I discuss one variant of a method for developing a theoretically-based framework for comparing problem solutions, and the possibilities and difficulties inherent in the process. Then, using the family problem example, I briefly examine another variant. Finally, I report the results of an attempt to have a group of experienced teachers apply the coding system developed in the family problem example.

There are two threads in the paper, not entirely separable: first, the development of a coding system which will reveal differences across individuals or groups in problem solutions; second, application of that system by teachers.

Background

The literature related to this study comes from several fields of research: first, storage in memory, in which interest has grown from memory for nonsense syllables to that for more complex phenomena (Kintsch, 1974); second, problem solving, in which interest has grown beyond simple problems with clear-cut solutions to ill-structured problems with complex solutions (Frederiksen, 1984); and third, assessment, in which it is increasingly recognized that the more complex goals of education lack appropriate assessment methods (Archbald and Newman, 1988). The summary that follows is an abbreviated version of a fuller discussion in Nagy (1990b).

Schema Theory

Researchers have posited entities named schemata (schema in the singular), whose role it is to act as organizing principles for complex memories. Schemata (Anderson, Spiro and Anderson, 1978) act as mental structures that incorporate general knowledge, and are more abstract than the particulars of a given situation. Interpretation of an individual's memories in terms of a schema involves matching elements in the data from the individual with generic slots or placeholders in the schema. For example, taken from the administrative study summarized below, in a slot labelled "running a staff meeting", different individuals will have different views or strategies to avoid confrontation, generate discussion, or build team participation.

Cognitive theorists have debated the psychological status of schemata. Abelson (1981) argues that schemata have psychological reality, rather than merely being organizers for the convenience of researchers. On the other hand, Alba and Hasher (1983) argue that stored memories are richer than the highly selected subset predicted by schema theory. Their perspective, that it might be appropriate to view schema theory as a method of imposing order on complexity, not necessarily involving any strong assumptions concerning the nature of human memory, is adapted for the present research.

As a device for imposing order and examining differences, schema theory holds promise. For example, Schallert (1982) notes that schemata become more elaborate and specific with experience. This suggests that an examination of the details of story-lines across individuals might be used to highlight differences in specificity or sophistication, differences which in turn might be linked to experience and/or expertise, as in the administrative problem discussed below, or to age or education, as in the family problem.

Schemata can be generated at varying levels of specificity. Returning to the example above, a schema placeholder at a very general level might be "staff relations". Within that level might be several more specific slots, including "running a staff meeting". Within the category, and even more specific, might be "avoiding confrontation", "generating discussion", or "building team participation". Then, within each of those categories might be several possibilities, themselves differing in quality. Differentiations can go on indefinitely until, in the extreme, every datum has its own category. The point at which such specificity produces useful views of the data depends on the purpose of the examination.

One of the many difficulties with such a perspective is how to decide what constitutes a more complete version of a schema, i.e., making judgments of quality. Horton and Mills (1984) concluded that the schema approach is plagued by the lack of an independent definition of depth of processing or sophistication. Thus, a present limitation to the technique is reliance on subjective decisions concerning the value (i.e., level of sophistication, worth) of particular pieces of data. The connection between memory research and problem solving consists in viewing solving of a problem as retrieving story elements from memory.

Problem Solving

The focus of this study is on "ill-structured" problems (Frederiksen, 1984), which are characterized by greater complexity, less definite criteria for deciding if a solution has been reached, lack of complete information, absence of a "legal move generator", and no convenient list of accepted procedures. They also have higher verbal content and are more context dependent. Most "real-life" problems would be classified as ill-structured. Voss and Post (1988) noted that the method chosen for the analysis of ill-structured problems reflects the theoretical concerns of the investigators. Three examples demonstrate the variety of theoretical concerns and approaches used.

Larkin (1980), primarily concerned with teaching, has worked in physics and algebra, areas which exhibit some characteristics of both well-structured and ill-structured problems. She has found that large-scale units such as Schank's (1974) scripts, similar to schemata, are useful in the analysis of problem solving in such domains. Voss, Greene, Post & Penner (1983) set for their subjects the problem of the lack of productivity of the Soviet agricultural system. Their main concern was understanding the problem solving process (Voss and Post, 1988). They categorized statements as goal statements, which deal with relatively global moves, such as identification of major issues and subproblems, or reasoning statements, which deal with the analysis within the structure of these subproblems. Finally, Lawrence (1988), also concerned with understanding of the problem solving process (Voss and Post, 1988), worked in the context of judicial decision making. Her basic model consists of elaborate if... then statements. She spends considerable effort on the need for an analysis system to capture a priori perspectives ("frames of reference"), which correspond, according to Voss and Post (1988), to the magistrates' courtroom schemata.

Voss and Post's (1988) linking of methodology to theoretical framework is germane. The motivating concern for the present study is to expand the arsenal of assessment devices available at the school level. Thus, when faced with the choice between richness of detail and operational simplification, the latter must be chosen.

Assessment

There is considerable dissatisfaction with the impact of traditional (i.e., multiple-choice) standardized testing programs on school curricula. Nagy, Traub and MacRury (1986), in a review of this literature, point out the danger that what is most easily assessed tends to become most important. At the same time, there is a movement toward the teaching of "higher-order" thinking (Resnick and Klopfer, 1989). Despite progress in assessment methods (e.g., Nickerson, 1989), there is some antagonism between the teaching and assessment of higher-order performance and traditional standardized testing. Calls for improvement in assessment (Haertel, 1986; Archbald and Newman, 1988; Stiggins, 1988) tend to be calls for development of technologies beyond the multiple-choice item. In addition, there is a wealth of evidence that much

theoretical progress on assessment of complex outcomes has not reached the teacher level. Many of the articles in Anderson (1990) document this problem. The intent of this paper is both to contribute to theoretical understanding of how to assess in ill- structured domains, and to examine how teachers deal with this task.

The Administrative Problem

Introduction

The following is an abbreviated report of a study reported in Nagy (1990a). The methodology, adapted from Voss, Greene, Post and Penner (1983), has developed over several related studies of how principals solve problems. Full details of the evolution of the methodology are available in Nagy, Allison, Allison, & Moorhead (1990).

The study was an analysis of the responses of 31 practising elementary school principals to a case study involving conflict between a school staff and the school librarian. The situation involves elements of supervision, curriculum, policy, interpersonal relations, physical plant, budget and supply, and staff attitudes. The situation was presented to the subject for solution in the role of a principal new to the school.

Subjects were trained in the think aloud process and were asked to read the case aloud, interjecting their thoughts as they read. Then they were asked to think aloud about how they would solve the problem, and finally to recall their thought processes. We aimed for eight subjects from each of four experience groups, Aspirant (0 years), Novice (1-2 years), Seasoned (10-15 years) and Veteran (20 plus years); due to equipment malfunction, we ended up with data from only seven Seasoned principals.

Analysis and Results

The essence of the procedure used to analyze the data was to build a collective story-line, across subjects, capturing the variety of responses to the problem, including values exhibited, perspectives taken, and actions planned. Within this collective framework, then, individual responses to the problem were highlighted and compared. Table 1 gives a simplified version of the schema built from the collective responses of the 31 subjects, along with the percentage of the statements that fell into each category. Briefly, subjects dealt with the problem largely as one involving the librarian, the library, and the staff (Categories 5, 6, and 7). They talked about the problem solving process itself (Category 1) largely because they were specifically asked to, and talked about context (Categories 2, 3, and 4) relatively little. The reasons for ordering the categories as in Table 1 are discussed below.

Table 1

Simplified Schema for Portrayal of Solutions to Administrative Problem

1. PROBLEM SOLVING (17%)
    - 1.1 Definition of the problem
    - 1.2 Problem solving process

2. COMMUNITY (1%)
    - 2.1 Seek community and student input

3. SYSTEM(2%)
    - 3.1 Ask about board policy ot procedure
    - 3.2 Consult with colleague
    - 3.3 Bring in library resource
    - 3.4 Bring in personnel resource

4. SCHOOL (3%)
    - 4.1 School goals
    - 4.2 Atmosphere of school
    - 4.3 Timetabling
    - 4.4 Vice-Principal

5. THE LIBRARY (25%)
    - 5.1 Role of the library
    - 5.2 Present practice
    - 5.3 Improvement

6. PERSONNEL (22%)
    - 6.1 Entry, data gathering and rapport
    - 6.2 Conflict, trust
    - 6.3 Staff meetings
    - 6.4 Staff development and supervision

7. LIBRARIAN (30%)
    - 7.1 Diagnosis
    - 7.2 Data collection
    - 7.3 Transfer resolution
    - 7.4 Support

The full category system is more detailed (see Nagy, 1990a). It contains 7 slots for data at the second decimal level, 20 at the third, 47 at the fourth, and 12 at the fifth. It is possible to, in effect, start at the top of the full category system

and read a story giving a summary of what everyone chose to discuss in their responses to the case. Such a story-line appears in the original report.

Using this system to organize and tabulate subjects' protocols, three types of results were found. First, it was possible to isolate differences related to experience. For example, while Aspirants and Novices were quick to make statements concerning the importance of building trust and avoiding conflict, only experienced principals gave us explicit strategies for doing so. In contrast (perhaps), those less experienced were more inclined to discuss the importance of provincial library policy in the case. Second, we presented summaries of actions taken by the subjects to a panel of three independent experts, professors of educational administration. These experts gave quite consistent ratings on a 10-point scale, with which we isolated the five most high rated (two Novices and three Seasoned) and the five most low rated (three Novices and two Veterans) solutions. Again, we found specific differences. For example, the high scoring subjects consistently solicited the personal feelings of the librarian and helped her to develop ownership of her problems. Finally, we were able to compare the differences between experienced and inexperienced people with those between high and low scoring people. We found some characteristics of highly rated solutions that correlated with experience, and others with youth, or at least lack of experience. For example, both those with high experience and high ratings talked of planning beforehand for staff meeting, and having information available. In contrast, those with high ratings and those with low experience talked of setting priorities (i.e., subproblems within the larger problem) and of providing professional development for the librarian in both library and interpersonal skills.

Commentary

Several conceptual and methodological problems which arose in the data analysis are central to the present discussion. If we are to regularize such a process for the systematic analysis of complex verbal or written output, then the procedures need to be simple, agreed upon, and consistent. If we are to further put the technique in teachers' hands, it must be time-efficient. It must also be rewarding to the individual teacher, in the sense of producing information useful for the instructional process. Given the realities of present student grading systems, it must also be in some sense quantifiable; this last point, however, is not touched upon in this paper.

The initially encountered issue was segmentation of the protocol into units for analysis. The goal is to isolate "thought units" from each other. However, determination of when one thought unit ends and another begins (in effect, when the subject changes topics) depends on striking a balance between capturing detail and producing a manageable category system. The problem is made more complex by the ill-structured nature of the field; the category system of necessity evolves as you work. Technically, by using a word processing system with an automatic paragraph numbering feature, it was possible to adjust the protocol segmentation to the evolving category system. However, an inherent difficulty

in the segmentation process concerns individual speaking and thinking styles. Some subjects spoke at length on a particular topic, staying within a single category. Others crossed between the analysis categories several times within the same length utterance. The choices faced by the analyst are to code a large verbal output as one unit, or to artificially segment it. If segmentation is chosen and segments labelled differently from each other, then this is accomplished only by introducing increasingly finer but less useful levels of detail into the category system. If segmentation is avoided, then speech units of vastly different lengths and complexity end up with the same code. Either choice results in both gains and losses.

The second issue was organization of the category system. Initially, an attempt was made to use "actions taken" as the organizing principle. There were, however, too many slightly different actions for this to be feasible. Such a system would have resulted in huge numbers of categories at the first level of specificity. Instead the principle used was major areas broken into sub-areas, followed by detailed actions within the sub-areas. As Table 1 displays, seven major areas fell out of the data. Many categorizations, even into these seven large areas, were difficult decisions. To minimize this problem, analysis proceeded from the more global to the more specific areas, that is, in the order in Table 1. First, statements on the problem solving process were labelled and categorized, then on the community, and finally on the librarian. It was a more precise and less taxing task to postpone categorization only when a statement clearly fit a smaller category than to postpone when the statement only fit a larger category. Once within the seven larger categories, development of the sub-category structure was a relatively straightforward task.

The third issue is that of level of detail. The analysis produced of differences across principals in approach to the case reveals some interesting differences related to experience and rated quality. Many of these offer food for thought. However, despite the fine-grained analysis, more detail from the transcripts would be welcome in numerous instances. Further work is required (and scheduled) to complete the analysis of the transcript data. To illustrate the type of analysis required, consider one element of the schema at a more detailed level than Table 1, Category 6.1.2, entry and familiarization strategies. In this initial analysis, 48 statements in this category were made by 20 different individuals. We have broken these statements into two smaller categories, 6.1.2.1, personal knowledge and 6.1.2.2, professional knowledge, but we have made no attempt to assess the variety or quality of treatments and suggestions within these breakdowns. To do this without producing an unmanageable coding system is a major undertaking.

In summary, the analysis is clearly successful. The collective schema approach to protocol analysis is manageable, even with very large data sets and very complex problems. The results show interpretable differences between inexperienced and experienced principals. As well, they can be related in a sensible manner to independent judgments of response quality. Further, it has been possible to identify elements of good problem solving which are related to

experience and others which those less experienced seem to possess. The instructional implications of such a result require much further work, but in principle it appears that the analysis is operating at an appropriate level of detail.

However, much remains to be done. We have no data on the reliability of either our category system or method of separating transcripts into segments. We do not know how or whether we ought to take into account the sequence of statements given by respondents. Perhaps most difficult to overcome, the reported analysis has consumed over 100 hours. If anything is to come of this that is both theoretically based and useful within the constraints of classroom life, more work is required.

The Family Problem

Introduction

The following is an abbreviated version of a study reported in more detail in Nagy (1990b). The purpose of the study was to explore an analysis of discussions, among groups of elementary school children, of a social problem. As in the administrative problem above, the analysis method used was an adaptation of schema theory set against a background of recent research on the solving of ill-structured problems. The particular data analyzed came from an ongoing and much larger curriculum project comparing methods of teaching thinking skills in the classroom. The context serves as a vehicle for discussion of methodological issues rather than a definitive test of the curricula under study: first, the data came from the end of year one of a three-year project; and second, the treatments were subject to a great many design vagaries.

The subjects came from eight rural elementary schools in three school board jurisdictions in Southwestern Ontario. Nine schools were part of the larger study; scheduling problems prevented data collection in one school. The nine schools were assigned to one of three treatments, and within each school one Grade 3 and one Grade 6 class were chosen for participation. To collect the discussion data, students were taken from class in groups of about five and asked to discuss the problem for ten minutes. Across the two grades and three treatments, 76 such discussions were recorded. For present purposes, details of the treatments are not required; they will be referred to as Experimental, Supported, and Control.

The focus of this study was an "ill-structured" problem. Groups of students were taken from class and presented, both orally and in writing, with the following situation (note that neither age nor gender is specified): "There are two children in the Puzzlewich family. One child is called Pat and the other is B.J. Both of the children receive the same allowance. Pat is involved in many after school activities such as music lessons, ringette, church choir, and youth group. B.J., however, just attends youth group once a week. Mrs. Puzzlewich is always asking B.J. to do extra chores around the house. She NEVER asks Pat to help out. B.J. complains to the mother that it is unfair to have to do all of the chores and yet receive the same allowance as Pat. 'I want an increase in

allowance.' The Mother says, 'You aren't paid for chores. Your allowance is just for being part of this family. You may not have an increase in allowance.' What do you think?"

Analysis and Results

Before outlining the analysis and results, a brief comment on the difference between this study and that involving the principals is in order. Development of a category system for the children's responses was a much more difficult and less satisfying task. There were two main problems. One was a general reluctance to talk; in some of the discussions the interviewer, who was instructed to remain as passive as possible, talked as much as all the children combined. Second, much of the discussion was unfocused; children offered ideas with little reference to previous speakers, and in a substantial proportion of cases, it was difficult to be certain of the point they were trying to make.

The methodology evolved during the analysis, eventually producing two products. The first was a method of tracking the degree of cohesion in the discussion — the extent to which it was a conversation among the group rather than five children taking turns talking to the one adult. Briefly, this was partially successful and promising, but will not be mentioned again. The second product was a two-level category system for the statements made, organized to reveal the basic collective schema with respect to family fairness. Unlike the principals' situation, it was not possible to create a story-line. Each category developed is a loosely held together collection of sub-categories that deal with roughly the same perspective on the issue.

The category development process is interesting. The analysis began with ad hoc development of an elaborate category system for the statements made. Once complete, it was cleaned by removal of several categories that served no purpose, usually due to lack of frequency. These included connectives, statements of facts and assumptions about the case, humour and fantasy, incoherence or self- contradiction, and comments on the progress of the discussion. Next, some very general categories were created to keep the data simple. These included prompts from the interviewer, general agreement, specific agreement, personal anecdotes, and details, which were usually expanding on a point beyond a level judged useful for the intended analysis.

At this stage, we had about 100 categories in five large groups: unfairness statements, proposed solutions, cautions about proposed solutions, comments on the relevance of age, and value positions. Several problems became apparent: the system was unwieldy, many of the categories were used only once or twice, many captured very subtle distinctions in meaning, and boundaries among the categories were unclear. Considerable collapsing and rearranging was both necessary and relatively easy. A second sorting, with some amalgamation and deletion, yielded six Position categories and seven Action categories as listed in Table 2.

Table 2

Category system for the family problem

(I) Position Statements

Position-1 — statements showing acceptance of responsibility for tasks, and awareness of the broader family context;

Position-2 — statements showing a disregard for responsibilities, including statements that Pat ought to help only when convenient;

Position-3 — statements that the family ought to operate on a monetary basis;

Position-4 — statements showing awareness of age, and its impact

Position-5 — statements about the feelings of anyone in the family;

Position-6 — statements that both chores and extracurricular activities have value for the individual engaged in them;

(II) Action Statements

Action-1 — solutions which involve differential allocation of allowance or balance between the story characters of activities, chores, and rewards;

Action-2 — solutions which involve achieving, by a variety of means, a balance between the story characters of activities, chores, and rewards;

Action-3 — weaker solutions involving fairness when convenient;

Action-4 — solutions involving unilateral action by B.J.;

Action-5 — solutions which involve emphasis on a process, such as discussion or keeping records, or setting up a schedule;

Action-6 — more responsible solutions involving family cooperation and sharing costs;

Action-7 — a catch-all for unlikely or irrelevant proposals.

Differences in quality of the responses are evident in the definitions of Table 2. Two paths for subsequent analysis seemed possible: one, categorize all discussions as enriched, typical, or impoverished, on the basis of all thirteen categories, and examine patterns across grades and treatments; or two, identify enriched, typical, and impoverished treatments on the basis of each category, and examine patterns across grades and treatments. Even a cursory examination of the data demonstrated that the first path would be impractical; the evidence across the thirteen categories was not consistent enough within a single discussion. Therefore, the latter path was chosen.

For initial purposes, the thirteen statement categories were subjectively rated for quality as follows:

	Position	Action
Typical	1	1, 2
Impoverished	2, 3	3, 7
Enriched	4, 5, 6	4, 5, 6

Approximately 72% of the Grade 3 responses and 61% of the Grade 6 responses were captured in the Typical Schemata. The corresponding figures for the Impoverished Schemata were 11% (Grade 3) and 13% (Grade 6); and for the Enriched Schemata, 17% (Grade 3) and 26% (Grade 6). One obvious difficulty arose concerning the subjectivity of the system. Two of the views dubbed Impoverished were more common among Grade 6 than Grade 3 students, and on reflection are best explained by increased self-centredness resulting from the approach of adolescence. The tentative and subjective nature of these categorizations ought not be overlooked.

Since Typical statements were so much more frequent than Impoverished or Enriched statements, different analyses were required. For the Typical statements, both frequency and variety within the three large categories were examined for each discussion. Note that the 13 categories in Table 2 were produced from more than 100 smaller categories. For the Enriched and Impoverished Categories, simple occurrence as a function of grade and treatment (Experimental, Supported, and Control) was examined.

There were differences across groups in the extent to which they expressed sentiments dubbed part of the Typical view of family fairness. These differences, however, did not fall into a simple pattern; given the circumstances, systematic differences were not expected. In the Enriched and Impoverished categories, grade differences were quite compelling. The Grade 6 students made proportionally more than twice as many of the Enriched Schemata statements, as would be expected, but, as mentioned, they also made more of the Impoverished Schemata statements. There were as well some discernible, although unsystematic, treatment effects.

Commentary

This analysis rests on some assumptions concerning the nature of the data which need to be discussed. In particular, we need to address the relationship between individual and group data. Schrag (1988) has argued that there is no

way of assessing the thinking required for a task unless we know what tools the thinker has available. Since thinking processes are not directly observable, they must be inferred from observation of the relationships between input and output, in this case between B.J. and Pat's family problem and the recorded discussions. This inferential difficulty is a commonplace; one needs to accept a reasonable amount of inference in cognitive research. In the present case, one needs to accept that generally the typical student's statements are an adequate representation of the typical student's thinking.

Perhaps more contentious is the acceptance that in particular each student's speech adequately represented his/her thought processes. That is, what effect did the group have on the ability of each individual to think out the issue and express an opinion? There is no evidence available on the question of whether some individuals felt compelled to either remain silent or voice passive agreement when faced with the expressed opinions of more assertive classmates. This is a real limitation to the data available. In the original report, the argument was put forward that with five students per group, a discussion length of 10 (that is, ten changes of speaker excluding the interviewer) means that the average student took advantage of the opportunity to speak twice. Seven of the 76 discussions were shorter than 10, while 57 were longer than 20. "While there are no data on individual behaviour within the group, it seems safe to conclude that, while some students might have been unduly reticent, a substantial majority probably took the opportunity to express their views" (Nagy, 1990b).

One difficulty of the method, identified in the literature (Horton and Mills, 1984), is that the categorization of statements is at root subjective. What one would like to consider as a deeper level of processing could as easily be construed simply as more like the sentiments adults would like to see children express. The problem with this feature of the method is demonstrated by the fact that the Grade 3 students appeared less selfish, in aspects of their protocols, than the Grade 6 students. One might choose to define growth in perception of the situation empirically by accepting what might be a natural outcome of adolescence. Or, one might choose to consider what is desirable from the adult perspective as a valid curricular goal, and take the Grade 6 results as undesirable. The problem remains, however, of having to distinguish level of moral development (however defined) from level of cognitive processing.

A second difficulty of the method is that, if we are going to examine very ill-structured problems with no inherently obvious "correct answers", the category system must evolve from the data. Where one starts in the data analysis is important. When to open a new category is an arbitrary decision, based on a subjective view of the history of the analysis. Whatever "category width" might mean in this context, effort needs to be spent in holding it somewhat constant. Tied in with the obvious issue of simple inter-rater reliability, already mentioned, it would seem important for different analysts to analyze the data in different orders. The issue is somewhat simplified by the possibility that an already-created category system, from an earlier investigation with different children, might be imposed on the data, but there still remains the difficulty of valid and reliable creation of that first set.

It is legitimate to ask what has been accomplished by an attempt to impose schema theory that might not have been done from a more traditional perspective, such as a relatively theory-free development of a "marking scheme". First, there is ample evidence that teachers require assistance in assessing higher-order outcomes of instruction (Haertel, 1986; Stiggins, 1988). Neither the identification of what constitutes higher-order thinking nor the development of appropriate marking systems is a trivial task. Both require development and imposition of a theoretical framework. Pursuit of notions of typical, impoverished and enriched story-lines for complex situations is appropriate to such a task. Second, there are calls from those studying the assessment practice of teachers (e.g., Stiggins and Bridgeford, 1985) for more focused methods. A method allowing comparison of practice with a well-developed image of what might be qualifies as a focused method, both for research and for the improvement of practice.

Teacher Application of a Category System

Introduction and Method

A small investigation was conducted to see how well a sample of teachers would be able to apply the category system in Table 2. Using a systematic search process through the data base of the children's transcribed discussions, 40 statements were selected. These included three statements initially coded as belonging to each of the 13 categories, plus an additional statement from the final Action category to produce an even number of statements. An instrument was put together consisting of the following: (a) an introduction and thank-you; (b) the original stimulus story and the 13 categories (plus a 14th "does not fit" category; (c) instructions with one explained example; (d) the 40 statements; and (e) a place for comments. For three of the 40 statements, excerpts from preceding discussion were included to provide more context for the task. This instrument was distributed to a sample of convenience of 10 elementary school teachers on one staff. They were asked to work individually, at home if they wished. Nine of the group, including one student teacher, returned the form.

Results

A comment is in order before reporting the results. When I began searching the data base to construct the questionnaire, I found myself questioning many of my own initial categorizations, and fighting the tendency to search for more clearcut examples. This reflects the difficulty of the categorization task.

The nine teachers reported requiring an average of 50 minutes to complete the task. Most reported the task to be difficult, and gave comments that fall into five general categories:

(a) difficulty in distinguishing positions from actions;
(b) insufficient context or information;
(c) too many categories with ambiguous definitions;
(d) difficulty in understanding what the student's point was;
(e) more than one point to the student statement.

Comments (a), (c) and (d) were certainly a problem in the initial categorization. Comment (b) could have been avoided by providing more previous statements as context. Comment (c) was somewhat disconcerting; while many of the original statements had been coded as dealing with more than one category, these examples had been deliberately excluded in the selection process.

The results were spectacularly disappointing. For only 7 of the 40 statements did a majority of the teachers agree with the original categorizations; in 11 cases, none agreed. Table 3a displays the nature of the disagreements. The most serious difficulties were with Action Categories 2, 3, 6, and 7. Position Categories 5 and 6 and Action-4 gave the least difficulty. In general, the most common problems were in changing position statements to actions statements, and in coding action statements as different actions.

Table 3

Summary of Agreements

I. With Initial Categorizations	
(a) Using 13 Categories	
Agreement with original	- - 29%
Disagreement within Positions	- - 5%
Code Position as Action	- - 19%
Disagreement within Actions	- - 28%
Code Action as Positions	- - 13%
Code as "does not fit"	- - 5%
(b) Using 3 Categories	
Agreement with original	- - 45%
Switching Typical and Impoverished	- - 14%
Switching Typical and Enriched	- - 19%
Switching Enriched and Impoverished	- - 15%
Code as "does not fit"	- - 5%
II. With Majority Opinion	
(c) Using 13 Categories	
Agreement with original	- - 45%
Disagreement within Positions	- - 4%
Code Position as Action	- - 12%
Disagreement within Actions	- - 20%
Code Action as Position	- - 13%
Code as "does not fit"	- - 6%
(d) Using 3 Categories	
Agreement with original	- - 58%
Switching Typical and Impoverished	- - 13%
Switching Enriched and Impoverished	- - 12%
Code as "does not fit"	- - 6%

There are less discouraging ways to examine the data. First, I collapsed the Positions and Actions into three broader categories each, Typical, Enriched, and Impoverished. This raised overall agreement from 29% to only 32%. Thus, little of the disagreement was between categories at the same level of perceived quality. Second, since a major problem was distinguishing actions from positions, (as one teacher stated, "'should' is not a good linguistic marker"), I collapsed this distinction, leaving only the three categories of Typical, Enriched, and Impoverished. This improved overall agreement (Table 3b) to 45%, but still left 15% of the responses with disagreement between the two extreme categories of Enriched and Impoverished.

The next step was to remove from the original categorizations their pre-eminent status, and to consider them as merely one of ten judgments. Adopting such a stance resulted in a change of category for 16 of the 40 items: one Position statement was recoded "does not fit", three Positions statements were change to Action and one Action to Position, and eleven Action statements were reassigned to other Action Categories. Category Action-6 disappeared entirely, absorbed into Action-1. As can be seen in Table 3c, this improved agreement considerably, from unacceptable to less unacceptable. When we group by Typical, Impoverished and Enriched, but still maintain the Position - Action distinction, agreement rises to 49%; when we abandon the Position - Action distinction, it becomes 58% (Table 3d). This would have to be classed as barely acceptable, but the problem of switches between the extremes of Enriched and Impoverished still remains at 12%.

Commentary

On an optimistic note, it can safely be assumed that errors of statement classification are randomly distributed. The data on frequency of occurrence of various statement types, as in Nagy (1990b), are likely pointing in the right direction, albeit with rather wide confidence bands. Category definitions can be tightened, more examples given, and training, or at least discussion of the instructions, provided. One teacher in the sample provided an alternative set of categories: awareness of others, self-interested, values work, shows responsibility, values money. Broader, more abstracted categories might be easier to apply with consistency than those used in this study.

The difficulty of lack of context is a fault of the questionnaire instrumentation, and would not be a problem in a situation involving full transcripts or live observation. Little can be done about the vagueness with which children express themselves, especially when the extent of this vagueness might be an object of investigation.

All of the difficulties encountered are exacerbated by contexts with less structure. Indeed, reflection on the differences between the two reported cases suggest that the term "ill-structured" might be too broad a category as work of this nature moves from the cognitive science laboratory into the classroom.

Discussion

One important difference between the principal performance on their problem and the children on theirs was that a multiple-level category system could be derived from the principals' data. This allowed development of a real framework for systematic comparison of individual responses. In contrast, the family situation did not have that structure. Children saw the situation as fair or unfair (mostly the latter), and offered various suggestions for changing it. These suggestions can be categorized in several ways, but mostly at a single level of detail. This appears to be as much a function of the task as the ages and sophistication of the samples.

The segmentation of protocols into units is at one level simply a methodological problem, but at another level, it lies close to the heart of the issues raised by Schrag (1988). If a subject makes an utterance of several statements' length, then presumably each idea is linked in some manner with the previous statements. This is an inference, but the degree of ease or confidence with which we make the inference depends on how the category system deals with the statements, and on how closely they lie in the transcript (i.e., whether they are separated by several statements on another issue). The system as developed for the principals' case does not deal with the linkages between categorized statements.

In retrospect, the category system that emerged from the family problem discussions was allowed to grow unchecked. The level of categorization attempted was too detailed, and was eventually abandoned in the inevitable collapsing of categories required to make the data manageable. This problem did not arise in the school problem. The decision on level of detail, taken fairly early in the development of the category system, held up as "just about right". Whether this too is a function of the topics and ages of subjects, or is more dependent on the experience of the analyst (the principal study was done second) is an open question.

The attempt to have teachers apply a category system was instructive if not entirely successful. The indications are fairly clear that a smaller number of more global categories would have been more workable. However, the point of this line of research is to develop a theoretical basis for grading complex educational outcomes. There is some unresolved tension between what will work in real situations and what can be grounded in theory. Such issues need to be resolved if the product of this line of research is to be more than the common sense of experienced teachers.

References

Abelson, R. P. (1981). Psychological status of the script concept. *American Psychologist*, 36, 715-729.

Alba, J. W., & Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, 93, 203-231.

Anderson, J. (1990) [Special issue]. *Alberta Journal of Educational Research*, 36(1).

Anderson, R.C., Spiro, R.J. & Anderson, M.C. (1978). Schemata as scaffolding for the representation of information in connected discourse. *American Educational Research Journal*, 15, 433-439.

Archbald, D.A., & Newmann, F.M. (1988) *Beyond standardized testing*. Reston VA: National Association of Secondary School Principals.

Frederiksen, N. (1984). Implications of cognitive theory for instruction in problem solving. *Review of Educational Research*, 54, 363-407.

Haertel, E. (1986). Choosing and using classroom tests: Teachers' perspectives on assessment. A paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Horton, D. L., & Mills, C. B. (1984). Human learning and memory. *Annual Review of Psychology*, 35.

Kintsch, W.A. (1974). *The representation of meaning in memory*. New York: John Wiley & Sons.

Larkin, J. H. (1980). Teaching problem solving in physics: The psychological laboratory and the practical classroom. In D. T. Tuma & F. Reif (Eds.), *Problem solving and education: Issues in teaching and research*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lawrence, J.A. (1988) Expertise on the bench: Modelling magistrates' judicial decision-making. In M.T.H. Chi, R. Glaser, and M.J. Farr (Eds.). *The nature of expertise*. Hillsdale, NJ: Lawrence Erlbaum Assoc.

Nagy, P. (1990a). Using schema theory to map ill-structured problem solving: An application to school administration. Presented at the Annual Meeting of the Canadian Society for the Study of Education, Victoria.

Nagy, P. (1990b). Assessing thinking skills in social problem solving. Presented at the Annual Meeting of the American Educational Research Association, Boston.

Nagy, P., Allison, D., Allison, P, & Moorhead, R. (1990). Perspectives on principals' problem solving. Presented at the Annual Meeting of the Canadian Society for the Study of Education, Victoria.

Nagy, P., Traub R. & MacRury, K. (1986). *Strategies for evaluating the impact of province-wide testing*. Toronto: Ministry of Education.

Nickerson, R.S. (1989). [Special Issue]. *Educational Researcher*, 18(9).

Resnick, L. & Klopfer, L. (Eds.). (1989). Toward the thinking curriculum: Current cognitive research. *ASCD yearbook.* Association for Supervision and Curriculum Development: Alexandria, VA.

Schallert, D.L. (1982). The significance of knowledge: A synthesis of research related to schema theory. In (W. Otto & S. White, Eds.). *Reading expository material.* Toronto: Academic Press.

Schrag, F. (1988). *Thinking in school and society.* New York: Routledge.

Stiggins, R.J. (1988). The nature and quality of teacher-developed classroom assessments. A paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Stiggins, R. J., & Bridgeford, N.J. (1985). The ecology of classroom assessment. *Journal of Educational Measurement*, 22 (4), 271-286.

Voss, J.F., Greene, T.R., Post, T.A., & Penner, B.C. (1983). Problem solving skill in the social sciences. In G.H. Bower (Ed.). *The psychology of learning and motivation: Advances in research theory.* (pp 165-213). New York: Academic Press.

Voss, J.F. & Post, T.A. (1988). On the solving of ill-structured problems. In M.T.H. Chi, R. Glaser, & M.J. Farr (Eds.). *The nature of expertise.* (pp. 261-285). Hillsdale, NJ: Lawrence Eribaum.

Construction of Curriculum Relevant Tests by Teachers and Experts

Bikkar S. Randhawa
University of Saskatchewan

Tests are an important part of the educational enterprise. Miller and Erickson (1985), in their introduction to a guide for planning, constructing, administering, and interpreting teacher-made tests, underscored the importance of student testing and the fact that proper attention was not paid to this aspect. They stated:

Probably no aspect of education is more talked about and less attended to than student testing. There are several reasons for both the insufficient attention and the poor test construction— time pressures, inadequate test construction skills, and incorrect judgment about students' ability levels. (p. 5)

It is mainly through testing, both formal and informal, that teachers can determine the status of the students in their care. Effective teaching requires that instruction be at the level at which students can benefit. How do we determine the optimal level of instructional focus? To address this question it must be realized that testing and evaluation are especially important since most instructional situations are such that a class, not an individual, of heterogeneous knowledge-base, abilities, motivations, and processing skills is the target of instructional intervention.

Before addressing the above question let us consider the components of instruction. Instruction is the process of presenting academic content in the form of knowledge, concepts, principles, generalizations, and applications for achieving the curricular goals. Instruction would entail analyzing the curriculum, an organizational plan for presentation, decisions regarding modes of delivery taking into account the entry behaviors of students, and assessment of student progress at various stages of instruction. Randhawa (1971) presented a view of instructional system specifically for individualization. This system incorporated an information processing unit, TOTE, first proposed by Miller, Galanter, and Pribram (1960). A brief outline of this system is given in Fig. 1.

Gagne and Briggs (1979) proposed a systems approach for instructional design. This system represents a series of fourteen stages that begins with analysis of needs and goals toward eventual demonstration that a proposed system of instruction is successful in meeting the stated curricular goals. The most important element in instruction is decision-making. It is through a series of decisions that an instructional plan is drawn and implemented. Decisions cannot be made in vacuum. The context of instructional decisions is varied and multi-faceted. Defensible decisions can only be made with reliable and valid information on students, resources, and facilities. Gathering such information in an efficient manner is the responsibility of instructional designers. Teachers become key players in the instructional design for their classes. Thus, it is imperative that teachers have the know-how for gathering the relevant informa-

tion for instructional decisions. The focus of this paper is on the construction, use, and effectiveness of teacher-made curriculum relevant tests. For comparative purposes teacher-made tests are compared and contrasted with standardized tests. Also, strategies for improving teacher-made tests are presented, and it is argued that these are highly dependent upon mandatory and improved instructional opportunities in the pre-service teacher education programs.

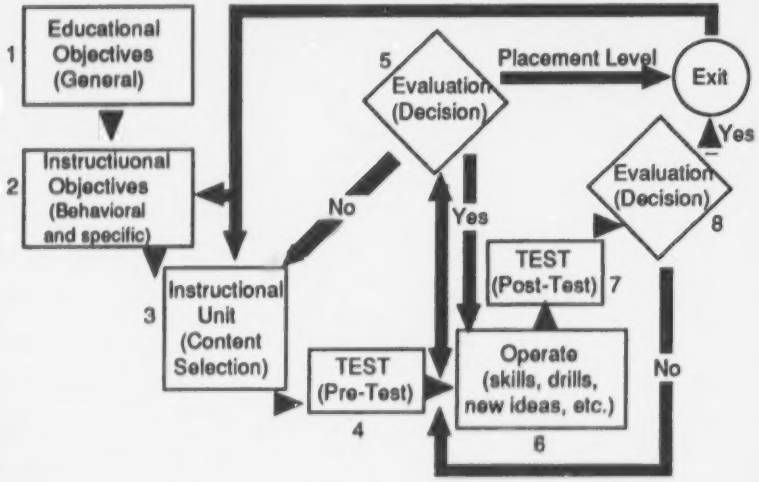


Figure 1. An instructional model for group or individualized instruction.

Teacher-made Tests

Extent of Testing

A typical teacher spends between 10 to 15% of the instructional time on the assessment of student progress and on diagnostic information gathering (Carlberg, 1981; Newman & Stallings, 1982). Gullickson (1982) found that 95% of the teachers he surveyed tested at least biweekly. A recent survey conducted in Alberta indicated that evaluation activities of a teacher took up approximately 25% of the instructional time allocated to a subject (W. T. Rogers, personal communication, May 10, 1990). Instruments used for this assessment by teachers are usually home-made. In contrast, the results of standardized tests, which are administered annually and at most grade levels in the U. S. schools, are used infrequently for instructional decisions by classroom teachers (Beck & Stetz, 1979; Fennessey; 1982; Stager & Green, 1984).

Surveys of teachers have indicated that on the average between 40 to 50% (with a range of 0 - 100%) of the course grades of students are dependent on test

scores (Gullickson, 1984b; Mckee & Manning-Curtis, 1982; Newman & Stallings, 1982).

Comparative Personnel Support

Teachers in Canadian schools do not have as much assistance and testing resources available as in most U.S. schools. Most U.S. school districts have professional measurement and evaluation personnel in their employ. It is only in a few larger school districts in our country that we have such personnel providing assistance and support to the teachers. This is not because teachers in Canadian schools have necessarily better pre-service preparation in measurement, evaluation, and statistics. Furthermore, the in-service training provided our teachers in these areas may not be taken as the reason for not having adequately trained resource personnel available at the school district level in many jurisdictions. The attitude seems to be that teachers will somehow manage to do the job of assessment and indeed they have. But the quality of the job done has not been questioned in many quarters. This may be because we in Canada are too polite or that the academic community is afraid of being caught in a vicious debate as to the responsibility for any perceived deficiencies in our teachers in this or any other area. I too will skirt this issue. However, I want to bring to your attention one important fact from my own faculty. Only in one of the four major pre-service teacher education programs in our college do students take three credit units in measurement and evaluation. In all the other programs students spend only about one-third of the time on measurement and evaluation in a three credit class on learning and instruction. Only about five percent of the pre-service teachers who graduate from our college would have taken an elective statistics class.

Teacher Knowledge of testing and Measurement

Fennessey (1982), Gullickson (1984), and Newman and Stallings (1982) suggested that teachers' knowledge of testing techniques and their skill in classroom testing practices was less than adequate. Since teachers spend a significant proportion of their instructional time on testing or gathering information for instructional decisions, sub-optimal practices may be detrimental to promoting excellence in learning and instruction. There seems to be consensus in research on the reasons behind this situation. Some of these are: inadequate pre-service and in-service education of teachers; negative attitudes toward testing on the part of administrators, teachers, and students; lack of motivation to learn and use appropriate measurement techniques; environmental context of many schools and districts; time constraints; choosing instructional options which do not rely much on measurement data; and so on.

The number of purposes for which teachers administered tests and the number of item types they employed were related to the knowledge of measurement and evaluation (Newman & Stallings, 1982). Teachers with higher level of professed competence tended to use tests for more and appropriate purposes and to use more item types. However, Fennessey (1982) found no relationship

between training, and patterns of test use or types of tests administered. Knowledge of testing and measurement principles and concepts was not related either to the amount of time teachers allocated to testing or to the percentage of tests teachers constructed themselves. However, Yeh (1980) found that training in testing and measurement was significantly related to the use, not necessarily proper use, of standardized test results. But, he found that training in testing and measurement of the group surveyed was somewhat limited. A recent study (Green & Stager, 1986-87) reported that the number of courses teachers had taken in testing and measurement had little relationship to the frequency of test use but was significantly related to the use of contemporary measurement practices. They also found male teachers to have more positive attitudes toward all aspects of testing than did females. Furthermore, teachers who considered their own test results to be of value and who viewed tests as generally effective and fair seemed to use tests more extensively. Teachers in general reported themselves to be comfortable with their knowledge of testing and use of tests. However, Green and Stager (1986-87), and others (Gullickson, 1984; Leiter, 1976; Newman & Stallings, 1982), pointed out a lack of sophistication, "possibly even a lack of competence, in testing techniques, particularly in statistical analysis of test results" (p. 53). Newman and Stallings (1982) found little change in competency in measurement and evaluation of teachers since Mayo's (1967) study.

Conant (1963) in his book entitled, "The Education of American Teachers", pointed out deficiencies in the preparation of teachers and made specific reference to the importance of testing and measurement for teachers in order to properly test and evaluate their students. This book drew a response from the National Council on Measurement in Education in the form of a sponsored symposium, "The Implications for NCME Policy of Conant's Book" (Mayo, 1964). Three papers delivered at the symposium were published in the first volume of the Journal of Educational Measurement.

Mayo's (1964) paper was one of the three delivered at the symposium at the NCME annual meeting in Chicago. This paper reported the results of a survey of teachers, principals, superintendents, college and university professors, and testing and research specialists. The survey solicited the ratings of these respondents to 70 competencies derived from the Outline of Needed Competencies from the NCME Committee on Pre-service Preparation of Teachers in Educational Measurement. A majority of the competencies were deemed to be important by these experts. Only two of the 70 statements describing competencies were rated to be "Of Little Importance" on the average. The rest were judged to be "Desirable" or "Essential" by a majority of the raters.

Mayo (1964) indicated that competencies rated by the judges comprised four content categories in measurement and evaluation: construction and evaluation of classroom tests; standardized tests; uses of measurement and evaluation; and statistical concepts. In fact, the textbooks in testing and measurement which were subsequently written attempted to cover these content categories and the underlying specific competencies. In fact, the contemporary measurement and

evaluation textbooks "still strongly reflect the stated emphases from that study and a subsequent one (Mayo, 1967)" (Gullickson, 1986, p. 347).

As noted earlier, teachers do not consider statistical analysis of tests or items important and they put more emphasis on non-test data in their evaluation of students. In order to determine the relative perspectives of teachers and professors on measurement and evaluation Gullickson (1986) conducted a questionnaire study of 24 professors and 360 teachers from elementary and secondary schools. The respective teacher and professor questionnaires presented a list of 67 topics from eight measurement content categories to be rated. The professors were asked to indicate the instructional emphasis they put on the topics in their pre-service measurement courses and the teachers were asked to indicate the relative emphasis they believe should be placed on each topic. Ratings on the topics were summed for each content category. The results of a multivariate test showed significant discrepancies between teachers and professors mean ratings for the eight content categories. However, on five of the content categories univariate results were reliable. The best agreement and the highest assigned priority by both groups was on "the preparation of examinations". The other two categories the groups agreed on were: administering and scoring tests; and general assessment information. The categories on which professors and teachers disagreed on the emphasis in pre-service education were: non-test evaluation activities; formative evaluation; summative evaluation; legal issues; and statistics. Professors' mean ratings were significantly higher on statistics and lower on the other four categories than those of the teachers'. This is not surprising. Beck and Stetz (1979) suggested that measurement specialists had relatively inaccurate perceptions of teacher testing behaviors and needs. The situation has not changed since then. But the problem is not of perceptions, whether accurate or inaccurate; it is to determine what is relevant and realistic for classroom teachers to acquire during pre-service education and in what they will aspire to develop further competence because it is important for their professional competence.

Test Use

Gullickson (1985) reported that elementary teachers rely for their student evaluations more on non-test data than test data. On the other hand, secondary teachers put more emphasis on test than non-test data for student evaluation purposes. He found that both elementary and secondary teachers do not put heavy emphasis on commercially prepared tests; however, secondary teachers put stronger emphasis on these than elementary teachers.

In an earlier study, Gullickson (1984b) sought to determine teachers' perspectives of their instructional use of tests. He reported that teachers believed that they were on their own as far as testing for instructional purposes was concerned. Furthermore, teachers "appear(ed) to be comfortable in their knowledge and use of tests, perhaps too comfortable" (p. 247). While teachers felt they were pressed for time, they preferred to give tests frequently and believed that their preference was shared by both administrators and students.

Gullickson also found that teachers believed that tests were helpful for instruction and for evaluation of student progress. Teachers also believed that tests had a potent affect beyond their prescribed role, viz., assessment. Teachers viewed tests more positively for measuring lower cognitive outcomes and virtually all the respondents, teachers, agreed that tests should not serve as the only basis for the determination of student progress or grades.

Measurement specialists, e.g., Gronlund (1981), emphasize that item analysis data provide a basis for useful class discussion of test results, for remedial work, for general improvement of class instruction, and improved skill in test construction. However, elementary and secondary teachers believe pre-service courses in measurement and evaluation should emphasize the topic of test preparation, but they do not consider statistical analysis of test results of such an importance to be relevant at the pre-service level (Gullickson, 1984a). Without resorting to statistical analysis, teachers believe their tests to possess sound psychometric properties (Farr & Griffin, 1973). Furthermore, teachers do not perceive item analysis as practical in the classroom setting (Gullickson, 1984b). Hence, "without systematic analysis of these tests, teachers do not have assurance that their tests function as desired. At best this means teachers realize less than the full potential of their tests. At worst, many tests may misdirect teachers and their students" (Gullickson & Ellwein, 1985, p. 17).

Practical Problems with Teacher-Made Tests

It should be clear that the errors or problems commonly found in teacher-made tests could be generalized to instructors at the post-secondary level. As mentioned earlier, the problems to be enumerated below stem from three basic reasons: time pressures; inadequate skills in testing and measurement; and incorrect estimate of students' ability levels (Miller & Erickson, 1985). The errors that teachers commonly make in testing students have been summarized by Ebel (1980) and the following list is derived from this source.

1. Teachers tend to rely primarily on their own subjective, but presumably absolute, standards in evaluating achievement.

2. Teachers tend to put off test preparation to the last minute, then they do it on a "catch-as-catch-can" basis.

3. Many teachers administer tests that are too poorly planned, too short, or too inefficient in form to sample adequately the intended content and abilities in the subject.

4. Teachers often put too much emphasis on trivial or unnecessary details in their tests but neglect to include basic principles, understandings, and applications of the subject.

5. Teachers often write test questions, both essay and objective, whose effectiveness is reduced by ambiguity or by irrelevant clues to the correct answer.

6. Many teachers underestimate or overlook the influence of sampling errors on test scores.

7. Most, if not all, teachers fail to examine the effectiveness of their tests by even a simple statistical analysis of the items or the results of their tests.

Besides the identified problems of teacher-made tests, another pernicious problem is the interpretation of scores from these tests. It is not uncommon for many to regard the score as the grade. A 70% score on a test is taken, in many instances, as the grade without taking the trouble to evaluate it in the context of the purpose for which the test was given. In instances where a number of tests are administered with varying weights, the scores are weighted without regard to the unit of measurement involved at each measurement occasion. These situations result in inappropriate testing emphases being given to the different sections of the subject being evaluated.

Expert-Made or Standardized Tests

Standardized tests are usually prepared by teams of experts. A typical team involved in the construction of a standardized test would consist of a measurement expert, subject area specialist (s), teachers or consultants of a subject, editors, and associates. The number and type of expertise represented on a team for developing a standardized test varies from test to test. This is because the resources available and the coverage intended by the test are not the same from test to test.

A standardized test is intended for use over a diverse range of ability of students and encompasses a variety of curricula across a large number of jurisdictions. Therefore, a standardized test usually covers the general and common knowledge domain among several curricula. Such tests are not as sensitive to different instructional content and processes emphasized during instruction by a teacher as are the teacher-made tests. Standardized tests emphasize only generic skills and knowledge components which are assumed to be expected at the specified age or grade levels.

For testing the common and generic knowledge domain a painstaking attention to detail is paid in the development and field trials of a standardized test. First of all, a detailed table of specifications or a blue print of the test is prepared. For doing this all relevant curricula are consulted and once the test specifications are prepared these are sent to a number of consultants for reaction. Second, items are written. Third, items are administered to small samples from the target population. Fourth, responses to the items are analyzed and those elements of the usable items are revised which have been identified to be ambiguous or non-functional. At this stage, unsalvageable items are discarded. Fifth, two or more preliminary forms of the test are assembled. Sixth, each of these forms are administered to small representative samples of the population. Seventh, results of these administrations are analyzed and instructions, items, item arrangement, etc. are revised as necessary. Eighth, each of these forms are administered to representative samples of reasonable size according to a well defined validation plan. Ninth, results of validation administration are summarized in a manual. Some tests provide extensive details in a supplementary manual for experts and for use beyond the classroom.

Of course, the details sketched above are for an ideal standardized test. Many standardized tests do not meet all of these requirements. In order that test

makers and consumers are provided with some guidelines, the National Council on Measurement in Education, the American Psychological Association, and the American Educational Research Association have produced jointly *Standards for Educational and Psychological Tests* (1985). In spite of these guidelines, many unscrupulous test makers publish standardized tests which do not meet even minimally these standards. Therefore, consumers ought to beware of the fact that the sheer availability of a test under the title of a standardized test does not necessarily make it a test that meets the standards. That is why the published and commercially available standardized tests are reviewed by experts and these reviews are found in many journals and in a comprehensive and authentic source, "Buros Mental Measurment Yearbook", which is published about every four years.

From the available array of standardized achievement tests it is possible to select a test which can serve a variety of uses. Among these are: instructional, guidance, administrative, and research. In the case of instructional uses, standardized tests have been found to be effective for the evaluation of learning outcomes, evaluation of teaching, evaluation of curriculum, learning diagnosis, and differential assignments within class. However, standardized tests may not be quite effective for grading and motivational purposes in instructional situations (Mehrens & Lehmann, 1987).

Standardized achievement tests can be effective for occupational and educational guidance. Among the variety of administrative uses of standardized achievement tests are for selection, classification, placement, public relations (informational), and curriculum planning and evaluation (Mehrens & Lehmann, 1987).

In contrast, the teacher-made tests are most effective for the evaluation of learning outcomes, grading, and motivation. In a study of 445 first-year teachers, Hall, Villeme, and Phillippy (1985) found that these teachers considered teacher-made tests most useful for their self-evaluation and for motivating student learning. These teachers, it seems, attached the greatest weight to state assessment results for judging the adequacy of teaching and the adequacy of instructional materials. Within each of the three test types, teacher-made, district-wide standardized tests, and state-wide minimum competency tests, the beginning teachers considered test results important for decisions regarding students' academic progress and for decisions about the diagnosis of student weaknesses.

Whereas one or more forms of a standardized test can be reused, the teacher-made tests are seldom appropriate for reuse as a totality, a few items may be reusable. While teacher-made tests are sensitive to instructional processes as well as to the taught curriculum domain, standardized tests cover only the common knowledge found in several curricula across the target population. Teacher-made tests allow inferences of student achievement only in terms of absolute standards, standardized tests allow inferences relative to other students.

Compromises and Solutions

Instructional and testing environments are progressively becoming more complex and challenging. Schools no longer cater to students from a mainly single cultural background. Classrooms no longer represent a narrow range of ability, aptitudes, and readiness levels. With the new curricular philosophy of a locally responsive and yet balanced curriculum, offering integrated instruction both in processes and content, demands on teachers' ingenuity and creativity have mounted. Students are expected, therefore, to acquire substantive or declarative knowledge (content) as well as procedural (how) and strategic (when or planning skills) knowledge. These changes to many may not signal a departure from what schools were doing or supposed to be doing. However, with explicit statement of objectives of schools encompassing three knowledge types, pressures for assessing them have been or would be direct.

As an example of the increasing complexity of the contemporary curriculum Saskatchewan is a case in point. Following an intensive study of various sectors of the Saskatchewan population, of the curricula in use, and of professional input, a Minister's Advisory Committee on Curriculum and Instruction Review issued a report entitled, "Directions", in 1984. This report provided a general framework for the future Saskatchewan curriculum. Following this report, a Core Curriculum Advisory Committee was appointed by the Minister of Education in the fall of 1984 to identify and recommend policies with regard to the core curriculum. This committee issued its report, "Program Policy Proposals", in January, 1986 which provided a framework for K to 12 curriculum in the province. The two major components of this curriculum, being implemented in Saskatchewan in a planned and systematic way, are the seven *required areas of study* and the six *common essential learnings*. The seven required areas of study are the conventional subject area groups: language arts; mathematics; science; social studies; health education; arts education; and physical education. The common essential learnings, a designation not usual in the literature but somewhat of an enigma on first sight, are nothing more than general skills, attitudes, values, and appreciations. These are also similar to various intelligences identified by Gardner (1983). Specifically, these general skills are: communication; numeracy; critical and creative thinking; technological literacy; independent learning; and personal and social values and skills. The common essential learnings are intended to be incorporated into the required areas or subjects in the school curriculum.

With this kind of curriculum emphasis, it is clear that teachers are required to test not only the three types of knowledge within each subject but also the generic skills, attitudes, and dispositions. When the testing demands of these contemporary curricula and the teacher preparation curricula of programs across the nation are juxtaposed, the obvious conclusion is that the training programs are not in keeping with the assessment proficiency needed by the present and the future teaching force. The first solution is to include more but appropriate testing and measurement components in the pre-service teacher education programs. Only the inclusion and mandating of such curriculum in

the pre-service teacher education program would not be enough. Those teaching the measurement and evaluation components or courses must use approaches which motivate students to seek further upgrading and study beyond what can be taught during the formal education phase.

Also, it should be recognized that these undergraduate students, if properly motivated and stimulated during their early exposure to measurement and evaluation, would be our future graduate students and colleagues. We need to bolster the measurement and evaluation expertise of the personnel at the school district and the provincial department of education levels. Only through advanced training in these areas can we hope that teachers will have ready access to the expertise for exploiting the future developments in computerized adaptive testing and the commercial item banks.

Another possibility is that curriculum guidelines for various subjects be grade- and age-specific and provide for use by teachers item banks linked to various instructional processes and contents such as those produced by the Instructional Objectives Exchange at UCLA under the direction of Jim Popham. This format provides a convenient access through a sophisticated identification of items desired for a measurement event. I understand that attempts at developing item banks in Ontario have been made but I am not quite aware of the extent and quality of use of this source. However, if computerized or hard-copy format item banks are developed, then a periodic evaluation of the effectiveness of utilization and satisfaction with it ought to be carried out. During the development of such banks, it is important that teachers who have taught a particular subject at a specific grade level be involved in the development of them. Alternatively, curriculum specialists who are in touch with teachers and have access to instructional activities of teachers be involved.

A pernicious problem often identified in the literature (e.g., Gullickson, 1984) is the teachers' lack of use of even rudimentary statistical analysis and their ability to interpret test results and item analysis. This problem is inherent in the lack of confidence and competence of many teachers in their use and manipulation of numerical information. This is symptomatic of a corollary problem of innumeracy among many of our adults in general (Paulos, 1988). The only solution to this problem is the general improvement of instruction in and attitude towards mathematics. Unless that happens we will continue to witness unwillingness on the part of many undergraduate and graduate students to benefit from and enrol in measurement, evaluation, and statistics classes and programs. If we make the measurement, evaluation, and statistics components compulsory, without ensuring that students are properly motivated to learn and without making these students realize the importance of the knowledge in this area for their teaching function, pre-service teachers may just go through the paces to satisfy the requirement for certification. Our interest, to improve assessment practices of teachers, would not be served by this approach. We have to seek interesting and innovative ways of teaching and motivating students in this domain. It will be ideal if students seek these components even if these are available as electives. That is a formidable challenge in the near future.

In-service has been used in many instances to bolster the skill and knowledge repertoire of practising professionals. It should be used only to bring forth newest and innovative approaches to the teaching function but not as a substitute to a pre-service education in an area as important as evaluation on which teachers spend a considerable amount of their contact time with the students.

Inservice should supplement and complement pre-service background in any important teaching function teachers are expected to perform. A well educated and informed teaching force should initiate in-service in their perceived need area where more than the basic background in the area is needed. Inservice by no means is a substitute for a formal course or program in an important knowledge domain. We are too often lured into doing in-service at times and to the clients quite inappropriate for it to be effective.

Conclusion

The paper surveyed the use and effectiveness of teacher-made tests. It has been pointed out that evaluation plays a prominent role in the teaching function of a classroom teacher. Teacher-made tests are the primary means of grading students and providing feedback to the students and teachers. In spite of this crucial role of teacher-made tests, many teachers do not receive adequate pre-service education in measurement, evaluation, and statistics. It is argued that education in this area be mandatory and effective.

Teacher-made tests have been compared and contrasted with the expert-made or standardized tests. Finally, a set of proposals for improving teacher competence in measurement and evaluation have been made.

References

Beck, M. D., & Stetz, F. P. (1979, April). Teacher opinions of standardized test use and usefulness. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Carlberg, C. (1981). *South Dakota study report*. Denver, CO: Midcontinent Regional Educational Laboratory.

Conant, J. B. (1963). *The education of American teachers*. New York: McGraw Hill.

Ebel, R. L. (1980). *Practical problems in educational measurement*. Lexington, MA: D. C. Health and Company.

Farr, R., & Griffin, M. (1973). Measurement gaps in teacher education. *Journal of Research and Development in Education*, 7, 19-28.

Fennessey, D. (1982, July). Primary teachers' assessment practices: Some implications for teacher training. Paper presented at the annual conference of the South Pacific Association for Teacher Education, Frankson, Victoria, Australia. (ERIC Document Reproduction Service No. ED 229 346).

Gagne, R. M., & Briggs, L. J. (1979). *Principles of instructional design* (2nd. ed.). New York: Holt, Rinehart, & Winston.

52

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences*. New York: Basic Books.

Green, K. E., & Stager, S. F. (1986-87). Testing: Coursework, attitudes, and practices. *Educational Research Quarterly*, 11 (2), 48-55.

Green, K. E., & Stager, S. F. (1986). Measuring attitudes of teachers toward testing. *Measurement and Evaluation in Counseling and Development*, 19, 141-150.

Gronlund, N. E. (1981). *Measurement and evaluation in teaching* (4th ed.). New York: MacMillan.

Gullickson, A. R. (1986). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement*, 23, 347-354.

Gullickson, A. R. (1985). Student evaluation techniques and their relationship to grade and curriculum. *Journal of Educational Research*, 79 (2), 96-100.

Gullickson, A. R. (1984a, April). *Matching teacher training with teacher needs in testing*. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans.

Gullickson, A. R. (1984b). Teacher perspectives of their instructional use of tests. *Journal of Educational Research*, 77 (4), 244-246.

Gullickson, A. R (1982). The practice of testing in elementary and secondary schools. Vermillion, SD: University of South Dakota. (ERIC Document Reproduction Service No. ED 229 391)

Gullickson, A. R., & Ellwein, M. C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. *Educational Measurement: Issues and Practice*, 4, 15-18.

Hall, B. W., Villeme, M. G., & Phillippy, S. W. (1985). How beginning teachers use test results in critical education decisions. *Educational Research Quarterly*, 9 (3), 12-18.

Leiter, K. C. W. (1976). Teachers' use of background knowledge to interpret test scores. *Sociology of Education*, 49, 59-65.

Mayo, S. T. (1967). Pre-service preparation of teachers in educational measurement (Contract No. OE 4-10-011). Chicago, IL: Loyola University.

Mayo, S. T. (1964). What experts think teachers ought to know about educational measurement. *Journal of Educational Measurement*, 1, 79-86.

McKee, B. G., & Manning-Curtis, C. (1982, March). Teacher-constructed classroom tests: The stepchild of measurement research. Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Mehrens, W. A., & Lehmann, I. J. (1987). *Using standardized tests in education*. New York: Longman.

Miller, G. A. , Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart, & Winston.

Miller, P. W., & Erickson, H. E. (1985). *Teacher-written student tests: A guide for planning, creating, administering, and assessing*. Washington, DC: National Education Association.

Newman, D. C., & Stallings, W. M. (1982, March). Teacher competency in classroom testing, measurement preparation, and classroom testing practices. Paper presented at the annual meeting of the American Educational Research Association, New York (ERIC Document Reproduction Service No. ED 220 491)

Paulos, J. A. (1988). *Innumeracy: Mathematical illiteracy and its consequences*. New York: Vintage Books.

Randhawa, B. S. (1971). Meaningful learning and individualized instruction in relation to TOTE unit for an evaluation paradigm. *Manitoba Journal of Education*, 7(1), 57-68.

Stager, S. F., S& Green, K. E. (1984). *Wyoming teachers' use of tests and attitudes toward classroom and standardized tests*. Laramie, WY: University of Wyoming, Department of Educational Foundations and Instructional Technology.

Standards for Educational and Psychological Testing (1985). Washington, DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Yeh, J. P. (1980). *A re-analysis of test-use data*. Los Angeles, CA:Center for the Study of Evaluation, University of California, Los Angeles. (ERIC Reproduction Service No. ED 205 590)

Considerations for the Implementation of an Ungraded Primary Program

Katherine A. MacRury
Ontario Institute for Studies in Education

In 1989, the Ministry of Education for the Province of British Columbia took the decision to implement an ungraded primary program for children in their first four years of formal schooling, on a province-wide basis. While a Ministry document entitled The Primary Program (1989) did present an overview of classroom practices for teachers in the ungraded setting, the Ministry did not release a set of policies that would have guided the implementation of the new program. Within this context, the present paper considers five issues that are related to the implementation of the ungraded primary program and that may require specification in provincial policy guidelines.

The five issues for the implementation of this ungraded primary program are:

1. Documenting the features of the ungraded program at the level of the school,
2. Designing the program within the context of teaching resources,
3. Developing student assessment policies with respect to monitoring progress and diagnosing developmental problems,
4. Providing equal opportunity for each child to progress, and
5. Facilitating the child's transition from the ungraded program to a graded program.

1: Documenting the program features at the school level. It cannot be assumed that the ungraded program should be implemented with the same features across schools or school districts, for reasons that could include (i) the perceptions of the concept of "ungraded primary program" held by decision-makers from different districts, (ii) the special needs of children within particular school catchment areas, and (iii) the resources that are allocated for school or district level implementation. Thus, documentation of the programs implemented by schools would be essential to providing an accurate description of the province-wide program and a meaningful understanding of program effectiveness.

By necessity, it would appear that implementation of a provincial program will have features that differ at the local (school) level. In the research on the effectiveness of the open education movement of the late 1960s and early 1970s, differences across sites were seen as problematic. Several studies which attempted to synthesize the research on the effectiveness of open education programs reported on the difficulty of comparing results of different studies because of the variability of the operational definition of "open education" with, for example, the concept of openness emphasized as a physical space in some studies and as an educational structure in others (e.g., Horwitz, 1979; Peterson, 1979; Marshall, 1981; Giaconia & Hedges, 1982). In fact, Marshall (1981)

suggested that the concept of open education be eliminated altogether, in favour of being more precise about the differences in the component features of the implemented programs (p. 181).

There were also differences in the way in which the features of open education programs were categorized across different studies; for example, Giaconia and Hedges (1982) compared the categories of features used by Traub, Weiss, Fisher and Musella (1972) and Walberg and Thomas (1972) to their own. Of their seven categories, Giaconia and Hedges considered four program features to be educational treatments: (i) role of the child in learning, (ii) diagnostic evaluation, (iii) materials to manipulate and (iv) individualized instruction, and the other three, administrative or organizational features: (v) multiage grouping of students, (vi) open space and (vii) team teaching.

For the ungraded primary program in British Columbia, it is not yet known to what extent the provincial guidelines will provide the flexibility for school principals and teachers need to adapt the program to meet the needs of the school or school district. However, it is anticipated that the documentation of program features at the school level will be necessary to accurately describe the program from a provincial perspective.

2. Designing the program within the context of teaching resources.

The allocation of teaching resources to the ungraded program may be a factor in the extent to which program features can be implemented. From recent reports of personal experiences with innovative primary programs, the fundamental resource that seems to have determined the success of these programs is the classroom teacher. (see e.g., Charlesworth, 1989; Oberlander, 1989; Rothenberg, 1989.) Two of these sources are quoted below because they illustrate from the insider's point of view the dependence of the innovative program on the training and commitment of teachers:

From his experience as a former teacher and researcher in open classrooms, Rothenberg wrote that "Teaching in an open classroom, even in the best of circumstances, is very demanding, perhaps far more so than in a traditional classroom. Effective teaching in an open classroom requires constant planning, continuous innovation, a sensitive system of monitoring students' performance, and well-developed skills in maintaining order without being authoritarian. Maintaining the energy and commitment to do all this well is difficult for trained and experienced teachers. It is impossible to say what the toll might be on teachers who are inexperienced or not well trained, or who face resistance from administrators, parents, or children" (1989, p. 78).

Charlesworth (1989), a primary school specialist, described the drawback to programs of continuous progress and multiage grouping in terms of the requisite skills of the teachers involved: "Individualized instruction and regular monitoring of progress are essential to its success. Teachers in this type of program must be skilled diagnosticians and planners... Many primary specialists, among them myself, believe that open education faded in this country not only because of the back-to-basics movement, but also because teachers were not in on the initial planning, were not thoroughly trained, and did not receive needed support once

they got started ... Continuous progress with multiage grouping demands exceptionally skillful and creative teaching in order to work successfully" (p. 10).

From these descriptions, teachers in successful open or ungraded programs must be skilled at planning, monitoring performance, being diagnosticians and maintaining order. The extent to which teachers are trained and experienced at these skills and techniques may determine the level at which a school can adopt the concept of the ungraded program. The teachers are also described as having a strong level of commitment on a continuing basis. To be successful over time, a program may also need to ensure that the school has a strong level of commitment to the teachers involved, providing a support system of resources that may include teacher assistants and computerized facilities to assist in the organization of student data for recording, synthesizing and reporting on student progress.

3. Developing student assessment policies.

From the draft 1989 document, evaluation of individual student progress is to be based on "multiple observations (p. 154)" and "a variety of observation techniques to ensure reliability and reduce the possibility of human error in judgment (p. 11.10)". The document did not discuss the various purposes an assessment can have; at least two assessment purposes, monitoring progress and diagnosing learning difficulties, might be considered by Ministry policy.

Restricting the monitoring of student progress to observational techniques raises issues about the validity and reliability of the interpretation of student reports. It is not known to what extent teachers make descriptive and/or evaluative observations with common standards, and the reliability of observational descriptions across teachers cannot be estimated without further study. When the teacher and students remain constant, it is easier to assume that descriptions over time are meaningful because there has been a common reference group. However, when the teacher or the students change, it is not clear that observations over time are reliable. Thus, the generalized assessment of a child's progress over time may be limited by the standards of reporting by the child's different teachers.

The importance of early diagnosis in the first years of schooling might be illustrated by research on reading in which poor performance in the early primary grades has been found to lead to poor performance in later years (see, e.g., Horn and Packard, 1985; Butler, Marsh, Sheppard and Sheppard, 1985; and Juel, 1988). On the basis of a four-year longitudinal study of 54 students in a low SES school, Juel made this foreboding observations:

"A vicious cycle seemed evident. Children who did not develop good word-recognition skill in first grade began to dislike reading and read considerably less than good readers, both in and out of school. They thus lost the avenue to develop vocabulary, concepts, ideas, and so on that is fostered by wide reading. This in turn may have contributed to the steadily widening gulf between the good and poor readers in reading comprehension and written stories." (Juel, 1988, p. 445)

If success in reading is vital from the first year of schooling, the earliest diagnoses of reading difficulties would seem to be an important purpose of student assessment. If observational techniques are the only accepted method of diagnosis, the underlying assumption is that every teacher will have the observational skills with which to making accurate diagnostic judgements. The draft 1989 document states that the use of standardized tests is inappropriate for primary children (p. 13); if this policy applies to diagnostic purposes of assessment, it might be similar to throwing the baby out with the bathwater. The potential uses of standardized tests may warrant further consideration by the Ministry, particularly if the early diagnosis and treatment of developmental problems are crucial to the child's later success and if standardized tests can assist the teacher in identifying these problems.

4. Providing equal opportunity for each child to progress.

The fourth issue for consideration is whether the ungraded program does provide all children with the equal opportunity to progress. If inequities can arise because of differences in the learning needs of individual students or special groups of students, should the primary program guidelines acknowledge that learning needs should be met by alternative teaching practices? As an example, consider the research which found that children from a low SES (socio-economic status) background benefit most from teaching practices that involve direct instruction.

One type of research study involved the re-analysis data from various Follow Through programs from the early 1970s in which approaches for teaching economically disadvantaged primary children were developed. A re-analysis of one such program led Gersten, Darch and Gleason (1988) to conclude that Direct Instruction was the most effective approach to teaching low-income students in kindergarten. Ciccelli (1983) summarized the characteristics of Direct Instruction in which the teacher is described as being the center of attention — "as dominant leader and central authority, establishes and enforces rules for group behaviour (p. 425-426)". This is in sharp contrast to Rothenberg's description of the teacher in the open classroom (which might apply to the teacher in the ungraded setting as well) — as both a director, organizing the environment to meet the needs of the students, and an instructor, leading discussions for small and large groups and occasionally teaching lessons to a large group (1989, p. 73-74).

Direction and structure were also found to be characteristic of effective elementary schools in low SES urban areas. For example, Clark, Lotto and McCarthy (1980) reviewed 97 studies on exceptional elementary urban schools in which successful schools had "clearly stated goals and objectives" and "structured learning environments" directed toward improving student achievement in reading and mathematics (p. 469); and Levine (1982) identified two approaches in successful inner-city elementary schools in Los Angeles and Chicago: (i) group-based, mastery-learning reading instruction, and (ii) curriculum alignment. One of the limitations to the generalizability of the effectiveness research is that outcome measures of success are typically based on the results

of standardized achievement tests which are likely to be more sensitive to measuring growth in schools where the goals have been narrowly defined and, possibly, easier to measure using the multiple choice test item format.

If it were assumed that the same instructional practices are not equitable for all groups of primary students, then the implementation guidelines may need to ensure that school principals and teachers have the flexibility to implement certain aspects of the ungraded program and still retain the particular structure and practices that are supportive of their students' learning needs. If such differences in learning needs are not considered and accomodated in the ungraded program, the inequities may become apparent when these children are placed in their first year of graded schooling.

5. Facilitating the child's transition to a graded program.

The Ministry of Education's decision in 1989 was to implement a program that would be ungraded for the first four years of formal schooling. It is assumed for this paper that, at the end of their fourth year, most children will exit from the ungraded program and enter a graded program. At this transition point, important decisions could be made that would affect the child's future school experience — for example, deciding to what extent the child can handle an achievement-oriented program and assigning the child to a level of graded program.

It is anticipated that transition to a graded program may prove to be particularly stressful for students, their parents and teachers. One determinent may be whether the teachers in the primary program will assume a role in easing the transition for students. For example, will teachers begin to administer to their Year Three and Four students the kinds of tests they will encounter in the graded Year Five (Grade 4)? One dilemma is that, if teachers do familiarize their students for what lies ahead, they are incorporating a graded dimension to the ungraded program. Another stess factor might be the measures that are used to assess the child's readiness for the graded program. At one extreme, each child could be assessed on his or her individual merits, by interpreting the teachers' observational accounts collected in the student's portfolio over the four year period; at the other extreme, student assessment could be based on standardized test results if a test battery such as the British Columbia Achievement Tests for Grade 4 were administered to all students in their fourth year of the ungraded program.

By the time the first cohort has completed the first four years of the ungraded primary program, it is conceivable that another ungraded program will be in place for years five and up. Even so, it is likely that this cohort, with each student proceeding at his or her own pace for four, six or eight years, will eventually encounter a graded program in which each student will be assessed relative to the rest of the cohort or according to norms set by the behaviour or performance of students outside the cohort. It is not evident how well-prepared this cohort will be to survive in a competitive learning environment.

Conclusion

It is suggested that five issues be considered in developing policies on the ungraded primary program throughout British Columbia.

1. Documenting the program features at the level of the school.

School-level features are likely to vary because of diversity in local perceptions, student needs, and available resources. Thus, accurate descriptions of programs at this level of implementation are necessary in order to have an overall perspective on the provincial program, determine features that are common and idiosyncratic across schools, and identify the explanatory variables in an assessment of program effectiveness. To label the primary program in general terms such as 'ungraded' as opposed to 'graded' or 'developmental' versus 'traditional' is to overlook the complex features that contribute to the unique implementations of the provincial program at the local level.

2. Designing the program within the context of teaching resources.

From recent accounts of authors with personal experience in innovative primary programs, the success of these programs is dependent upon the skills and commitment of the teachers who are involved. At the school level, consideration of the teachers' skills and opportunities for professional skill development might lead to a more feasible implementation plan. Also, teachers should not be expected to provide all of the commitment; the school and school district should be equally commited to their teachers in terms of providing assistance such as computerized resources to assist with monitoring and reporting student progress.

3. Developing student assessment policies.

Regarding teacher assessment practices, the Ministry's original 1989 document emphasized the use of a variety of observational techniques and denounced the use of standardized achievement tests. Missing from the document was a discussion of the range of purposes of assessment and the relationship of assessment to teaching practices in the ungraded program. Two major purposes of assessment, monitoring student progress and diagnosing learning difficulties, need to be elaborated.

4. Providing equal opportunity for each child to progress.

It is not clear how the ungraded program can provide equitable learning experiences for every child if, as research has indicated, some children may benefit from having more structured instruction than other children. Acknowledgement and support of individual or group differences would require a very flexible implementation policy. Unfortunately, the potential negative effects of the ungraded program on children who are believed to benefit from increased structure will not be known until after those effects have developed.

5. Facilitating the child's transition to a graded program.

The transition from an ungraded to a graded program may be the point at which important decisions are made regarding the child's future direction in school. Two issues related to transition are: (i) whether features of the graded program will gradually be introduced in the years prior to the child's exit from the ungraded program and (ii) determining the method(s) of assessment to be

used to determine a child's readiness for and placement in the first year of the graded program.

A final comment:

The five issues for consideration raise these broader questions about the implementation of the ungraded primary program in British Columbia:

1. To what extent will schools and school districts have the control to adapt the provincial ungraded primary program to meet local needs and resources?

2. Does the Ministry of Education acknowledge and support the variations in implemented programs to meet the special needs of individual students?

3. What measures can now be put in place to monitor the implementation of the primary program?

4. What are the necessary components of a research program that would develop valid, reliable and feasible methods for the assessment of individual students?

References

Bredekamp, S. (Ed.). (1987). *Developmentally appropriate practice in early childhood programs serving children from birth through age 8 (exp. ed.)*. Washington, DC: NAEYC.

British Columbia Ministry of Education. (1989). The primary program. Victoria, B.C.: Government Printer.

Butler, S.R., Marsh, H.W., Sheppard, M.J., & Sheppard, J.L. (1985). Seven-year longitudinal study of the early prediction of reading achievement. *Journal of Educational Psychology*, 77 (3), 349-361.

Charlesworth, R. (1989). "Behind" before they start? Deciding how to deal with the risk of kindergarten failure. *Young Children*, 44 (3), 5-13.

Clark, D.L., Lotto, L.S., & McCarthy, M.M. (1980). Factors associated with success in urban elementary schools. *Phi Delta Kappan*, 61, 467-470.

Gersten, R., Darch, C., & Gleason, M. (1988). Effectiveness of a direct instruction academic kindergarten for low-income students. *The Elementary School Journal*, 89 (2), 227-240.

Giaconia, R.M., & Hedges, L.V. (1982). Identifying features of effective open education. *Review of Educational Research*, 52 (4), 579-602.

Horn, W.F., & Packard, T. (1985). Early identification of learning problems: A meta-analysis. *Journal of Educational Psychology*, 77 (5), 597-607.

Horwitz, R.A. (1979). Psychological effects of the open classroom. *Review of Educational Research*, 49 (1), 71-86.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, 80 (4), 437-447.

Levine, D.U. (1982). Successful approaches for improving academic achievement in inner-city elementary schools. *Phi Delta Kappan*, 63, 523-526.

Marshall, H.H. (1981). Open classrooms: has the term outlived its usefulness? *Review of Educational Research*, 51 (2), 181-192.

Miller, J.W., Ellsworth, R., & Howell, J. (1986). Public elementary schools which deviate from the traditional SES-achievement relationship. *Educational Research Quarterly*, 10 (3),31-50.

National Association for the Education of Young Children. (1988). Position statement on standardized testing of young children 3 through 8 years of age. *Young Children*, 43 (3), 42-47.

National Association of State Boards of Education. (1988). Right from the start. The report of the NASBE Task Force on Early Childhood Education. Alexandria, VA: NASBE.

Oberlander, T.M. (1989). A nongraded, multi-aged program that works. *Principal*, 68, 29-30.

Rothenberg, J. (1989). The open classroom reconsidered. *The Elementary School Journal*, 90 (1), 69-86.

Traub, R.E., Weiss, J., Fisher, C.W., & Musella, D. (1972). Closure on openness: Describing and quantifying open education. *Interchange*, 3, 69-84.

Walberg, H.J., & Thomas, S.C. (1972). Open education: an operational definition and validation in Great Britain and United States. *American Educational Research Journal*, 9, 197-208.

What Should a Classroom Testing Program Look Like? The Functional Factors of an Assessment Program in Primary Classrooms

John O. Anderson and Dan G. Bachor
University of Victoria

The predominant paradigm in educational measurement literature is based on standardized paper-pencil tests (Stiggins, Conklin & Bridgeford, 1986). Classroom assessment procedures are often viewed as analogues of standardized tests in that the assessment procedures used by teachers can be treated as single instances of test administration, discrete from other kinds of classroom activities. The manner in which classroom assessment is to be completed has been the focus of debate for a considerable period of time (Bachor, 1978, 1979a, 1990). For example, measurement specialists have struggled with the norm-referenced and criterion-referenced interpretive frameworks (Ebel, 1978; Popham, 1978), standards for testing (Educational Testing Service, 1987), how to modify information collection procedures to capture the potential to learn (Bachor, 1979b; Feurstein, Rand & Hoffman, 1979), the use of technology in data capture (Colburn & McLeod, 1983), and the frequency and scope of measurement (Fuchs, Deno & Mirkin, 1984). However, the suitability of this paradigm itself merits attention.

Teachers have to measure the extent to which instructional intent has been achieved by each student in the class. This task is complex particularly in the dynamic environment of the classroom (Anderson, 1990). Considered in this paper are two broad issues: what does classroom assessment look like and what should it be? To address the first issue we describe some findings of a study in progress investigating the classroom assessment practices in primary classrooms in two schools. The second issue is addressed more speculatively by offering reflections on what characteristics an assessment program should have given the context of the primary classrooms previously described.

Classroom Assessment in the Primary School

The classrooms considered in this paper were in elementary schools which are beginning to implement what is known in British Columbia as the Primary Program. The Primary Program document (Ministry of Education, 1989) is a draft description of the educational program to be implemented in the first four years of public schooling in British Columbia. This initiative was developed in response to The Report of the Royal Commission on Education (Sullivan, 1988). The B.C. Ministry of Education has expressed a definite view on the primary classroom and the assessment practices to be implemented within it:

We envision that classroom assessment will be continuous, formative, and reflective of what children can do, and that assessment techniques will mirror

instructional practices, thus bringing together curriculum and assessment into one seamless process. (B.C. Ministry of Education, 1988)

In addition, the emphasis of the Primary Program is on a wide variety of goal areas: intellectual, physical, aesthetic, social and emotional development. Each student is to be evaluated in relation to each of these goals.

The guidelines for the implementation of the new Primary Program in British Columbia schools, provide an overview of the assessment objectives to be achieved in classrooms and examples of how this might be done in terms of data collection procedures (anecdotal records, work samples, and observation guides). Implicit in these guidelines is a call for the implementation of a substantial, complex assessment program individually crafted for each primary classroom in the province and for each student in the classroom. However, the programmatic aspects of assessment are not addressed, more specifically, the following question is not answered: how do the various components operationally fit together to result in a functional assessment program of data collection, collation, interpretation and reporting?

Purposes and Procedures

In 1989, we initiated a study into the classroom assessment practices of teachers who were currently implementing the new Primary Program. These teachers worked in schools that had been designated "lead schools," as they were among the first in the province to introduce the Primary Program on a voluntary basis. Thus, we approached the staffs of these two schools with the request for eight experienced teachers who would be willing to collaborate with us in examining their current assessment practices. These primary teachers agreed to cooperate with us in describing their assessment practices conceptually and operationally, allowing their classes to be observed, and critically reviewing the descriptions and generalizations developed by the authors. The intent was to describe practices that are typical of those used in the assessment of student attainment. In addition, four teachers from a third "lead school" have been interviewed to further verify the generality of the procedures observed and described in the first two schools. The three schools are located in two school districts on Vancouver Island. The four participating teachers from each school had parallel teaching assignments: one regular teacher of early primary program, one regular teacher of older primary children, one music teacher, and one learning assistance teacher.

Although this study is still in progress, the operational characteristics and procedures of student assessment in the classrooms we are investigating can be described.

What Classroom Assessment Looks Like in the Primary Program

The assessment of student progress is a characteristic element of the primary classroom. Assessment is essentially on-going part of the school activity. Although the assessment activities of the teacher are not hidden from view in the classroom, the students are not made particularly conscious of assessment as a

separate element of the classroom. Rather evaluation takes place as "informal" collections of information about student performance in the classroom context. The only exceptions are in the case of "at risk" students where additional steps may be taken that include discrete testing or assessment such as paper-pencil tests of various types.

The general operational characteristics of classroom assessment of the primary schools involved in this study are tabulated below:

Assessment is a normal component of classroom activities, it is not usually a discrete activity - it is not separate from other classroom activities, nor is it a one-time event.

Assessment can be discrete when the purpose is to determine or confirm that a child needs additional assistance or special programming.

Constant information flow - the teacher is continually collecting information on student performance that can be related to the educational goals. There is some decrease in the amount of data gathering subsequent to a report card being issued.

Group testing is not part of the assessment procedure in classrooms. Assessment procedures are based on individual student observation and product evaluation.

Assessment is conducted in a wide variety of educational goal areas: intellectual, physical, social, emotional and aesthetic development.

Assessment activities are targeted on individual students and on particular aspects of the instructional program (tasks and performances that are expected of the student). The instructional targets might be conceived of as the path(s) students are to follow in attainment of the goals. This path is more familiar for academic areas (generally related to the goal of intellectual development), and so the academic targets for assessment purposes are better articulated. Although assessment requires that targets be identified and student status in relation to these targets be reported, in the view of teachers, this is not to suggest that there are discrete end-points for a student to achieve and then terminate further learning and development. There is an implicit open-endedness to learning so the development of the child is open-ended in the sense of continual, unbounded progress.

The purposes of assessment are twofold: to direct instruction and to provide solid information to report on student progress.

Assessment is used to identify outliers in relation to two aspects of the classroom: i) the performance of individual students; and ii) the goals of education. In the first case, students are identified who are either exceeding or not meeting one or more educational goals. However, this identification is not simply whether or not a child is achieving goals, but also whether the child is performing up to ability, whether he or she is expending appropriate effort, and the extent to which progress is being made. In the second case, teachers examine the general progress made by all students in meeting various educational goals to determine whether all the goals are being given reasonable and fair coverage.

Assessment is multi-modal in that achievement can be demonstrated in a number of different ways. For example, for the same goal area one student may

choose to write a report which is a relevant demonstration of goal attainment, another student may deal with the information through another medium such as a play or visual art. All modes could be considered equivalent, or at least sufficient, for the demonstration attainment of the same goal.

The procedures used generate descriptive information, generally of a textual nature. The information is reduced in that summarizations and notes are made by the teacher as she is recording the descriptions. In spite of this reduction there are large volumes of information collected for each student.

Procedures: Information collection

1. Observation is probably the most widely and frequently employed procedure for the collection of information for student assessment. The observation generally occurs as the students are working on projects, assignments or other school-related tasks. The focus of observation is upon emergent skills, abilities, work habits, emotional development, and appropriate social interaction (cooperation, for example). The use of observation in the assessment of student achievement assumes an internalized knowledge of age/grade expectations by the teacher.

Two different types of observation are conducted, in part, varying as a function of age of the children being instructed. First, teachers observe activities as they are displayed by various children in class. The targets of observation are the ongoing, normal classroom activities of the students. Thus, a record of emerging skills, abilities, behaviour, patterns of social interaction, et cetera is obtained. Recognizing that an example of "new" behaviour has occurred serves as a prompt for the second type of observation. Second, teachers select a single child or group of children prior to the beginning of a class to look for specific patterns. In this case, the observation techniques are regularized. Targeted children are observed about twice a day for a period of approximately 15 minutes, so that each child in the class is observed within a two day span. Teachers with younger children (approximately ages 5 to 6) tend to use both procedures, those instructing older children (approximately ages 7 to 9) tend to use only the latter procedure.

The main form of observation is periodic, targeted observation of a child during regular classroom activities and recording salient features of the situation. Observation is conducted using one of two major perspectives: First, it is used to monitor students as they carry out daily classroom activities associated with the primary goals. The purposes are to note the emergence of a first occurrence of a behaviour relevant to educational goals, something that a child has just displayed in class (may be able to do) and to note what a child is able to do (can do). Second, specific children may be targeted when there is some doubt that a goal has been met and where it is quite evident a child is having difficulty (can't do).

Another form of observation could be termed "structured" observation - the observation of the classroom elements the student interacts with and the static features of the classroom which provide context for interpreting observation of student performance. Examples of the targets of this type of observation include

the peers the students chooses to work with, the books selected for recreational reading or for researching a topic, the area of the classroom or school the student prefers to work or to play. The information collected in this fashion provides context for the interpretation of other assessment data.

2. Collection of work samples: approximately 1 or 2 each month per child, placed in portfolios or individual student file folders and retained as permanent records. The work samples are evaluated by comparing student work to the goals (intentions) of instruction.

The work sampled is predominantly written pieces, such as special projects on animals or journal pages; although art work, math problems, audio tapes of reading, and video tapes of characteristic behaviour patterns are also included. The work can vary one child to the next in the sense that for a given task (such as reporting on a story read), one child may complete a written report whereas another may generate a visual (art) display.

3. Checklists: are essentially a listing of the performances or tasks related to an intended outcome that a child is able to do. Used to note the presence or absence of skill development. Checklists tend to be used primarily by regular classroom teachers. In physical education, a checklist of physical skills is used to describe the skills a student is able to do - these include both physical and social skills, for example patterns of movement. Checklists are also used in other areas such as noting the accuracy of rhythmic clapping patterns or the ability sequencing activities. They are used to assist in the monitoring of attitude and effort expended by students.

4. Interviewing: asking students questions on an individual basis on topics of relevance to teaching - e.g.: what books a student chooses and why, in order to estimate ability and interest of the student. Provides some perspective on the student's view of activities and progress and also provides a focus of the processes a student uses in accomplishing tasks.

5. Student self-evaluation: This can take a variety of forms. Once or twice a month students are requested to provide (write or state during an interview) a description of their own strengths, their experiences and ability of working with other students. These are retained in the individual student folders. Student self-evaluation can also include students declaring when they are ready to read a particular text or students verbally estimating progress in a particular area. This is an area of considerable interest to a number of participating teachers. They want to promote the use of self-evaluation by students to the extent that it becomes the main form of student evaluation. The procedures would include the reporting of the student's own progress (and presumably lack of progress) to the parents. The processes of high-order cognitive processing involved in self-evaluation are viewed as being high priority outcomes related to the goal of intellectual development.

6. Editing common writing samples: each student is given a sample of a student's writing, and then each student in the class is asked to read and edit the work. A copy of each student's editing is retained about one piece each month.

Procedures: Information reduction or selection to yield reports.

The reduction and interpretation of information collected is conducted in order to develop reports on each student's progress. The reports are written descriptions (ranging from approximately 3 to 6 pages) of student abilities and skills accompanied by comments on student status in relation to the goals of education and progress towards these goals. The reports do not contain lettergrades nor comparative descriptions in the normative sense. The reports are essentially descriptions of representative current performance. Assessment information is selected that has been determined as being typical of a youngster's activities (processes and products) for any reporting period. The information is derived from the implementation of the processes described above.

The amount of information collected is considerable, so, selection and reduction has to take place. The goals of the Primary Program are used to guide the selection of information. A child's progress within each of the goals is the main focus. Descriptions of development within a goal area are supported by information collected through the processes described above. For example, written work samples will be selected from journals that reflect changes in skills or abilities, stories or other written products in which a child demonstrated a change from invented to conventional spelling.

The student abilities and skills are described in terms of what the child can do. The term "can do" is used frequently in the Primary Program and by those involved with it. This emphasis on "can do" can also lead to some problems when there is some concern on the part of the teacher over the child's performance in some areas to which attention should be paid (to an extent, "the cannot do's"). To an extent, reporting of problem areas appears to go somewhat against-the-grain in regard to the ethic of the Primary Program.

The writing of the reports revealed a difficulty in condensing all the information available on each child. Participating teachers pointed out that one has to be aware that it is "noticeable" events that are getting observed and generally these events will involve only certain students, so, conscious attention must be paid to the "quiet kids" in order to collect representative samples of information. This holds not only for student coverage, but also for goals - some goals involve more noticeable, accessible and understandable student performance (for example, student performance in reading can be viewed as more accessible than that related to emotional development). Further, in keeping with the emphasis on describing what the student can do, the comments could not be simply a non-committal "satisfactory" or "not up to expectation", but rather more descriptive of student status supported by collected information.

Questions regarding practice observed

Our research is still in progress and many questions remain unanswered (more are yet to be asked and many more yet to be conceived). Many of the questions are directly related to fundamental aspects of the assessment of student achievement. We are currently pursuing these questions with our

teacher collaborators but have not yet established sufficient resolution to report. Before presenting our speculations on the desirable traits of a classroom assessment program, a sampling of some of the gaps in our knowledge base may provide some useful context for considering these traits. Some of these questions are:

a. How are the instructional elements (the goals, learning outcomes, tasks, objective indicators and the like) identified? How does the teacher know what to focus on in observing students?

b. When conducting observations, how does the teacher determine which students to observe?

c. What concerns exist regarding the sampling of information and students? Is the same kind of information collected on each student for each goal?

d. How much information is generated and collected for each student?

e. To what extent are student behaviours comparable in the sense that the students could be responding to or interacting with different components of their educational environment yet the teacher could be viewing the behaviour as of "the same sort". In actuality one student could be exhibiting reading-type behaviour, another student" social" behaviour, and the teacher could be viewing both students as displaying aesthetic behaviour.

f. How representative of the child's repertoire are the performances collected?

The above questions are fundamental to assessment of achievement of educational goals regardless of the procedures used to collect, collate, interpret and report assessment information. The characteristics presented below are, to an extent, the answers we hope to find to these questions.

Assessment Programs in Primary Classrooms: What Should be Considered.

Assessment in the primary classroom, as described above, is continuous, individual and integrated into classroom activities. The procedures seldom take the form of responsive, group testing, nor do they yield data in a form that is easily manipulated by humans or machines. The assessment program of primary classrooms should have the following characteristics:

1. Relevant

The assessment program should be relevant in that the assessment itself should have a purpose that serves the attainment of the goals of schooling. The procedures used should be directly related to instructional activities - essentially what is termed by some as authentic measurement (McLean, 1990).

2. Fair

The assessment should insure that every student has a good opportunity to display target behaviours which are indicative of goal attainment. Contextual constraints such as test anxiety, punitive overtones, and unpreparedness should be eliminated. To the extent possible, students should see the assessment procedures as fair.

3. Accurate

The assessment should result in accurate information, in that the data are derived from representative sampling of the indicators of the goals of education

being assessed, and of the behaviours of the students being assessed. The indicators are to be solid representations of goal attainment.

The reporting should be based on interpretation that is founded on a solid information base. The information upon which evaluations are based should be available and interpretable to stakeholders - the parents, the students (to the extent feasible) and teachers.

4. Feasible

The procedures have to be feasible within the context of the primary classroom. The whole set of procedures must be viewed as "do-able" by users, and must maintain the characteristics of relevance, fairness, and accuracy.

5. Systematic

The procedures should be used systematically in the sense that each student is evaluated on the same basis as any other student. In other words, given the same underlying level(s) of goal attainment, the same kinds of evaluation will be produced.

There should be consistency of procedures for information collection, interpretation and reporting from one educational area to another; and from one student to another.

Each area of schooling should be evaluated with similar standards of information collection and interpretation. For example, the quality and consistency of information collected to evaluate reading attainment should be similar to the quality and consistency of the information collected to evaluate attainment in mathematics, or the goal area of physical development.

6. Condensable

The information collected has to be of a nature that procedures are available to reduce it to a communicable size without a significant reduction or change in meaning.

Summary

A paradigm of classroom assessment based on standardized tests does not fit the classrooms involved in this study. The information collection conducted by the primary teachers participating in this study was not a classroom activity discrete from others. The information collected was not based on student responses to test items but rather to more global elements of the educational environment such as teacher requests for particular task completion, general social interaction or student activity directed towards a particular student initiated project. To an extent, this lack of fit is unfortunate since there are many excellent data analysis procedures available for those assessment data that do fit, and there is a dearth of systematic procedures for the classroom assessment we observed.

One path to resolve this incompatibility is to change what is going on in the classrooms - make the data fit. This is not an acceptable route given the philosophy of the Primary Program. Another path is to disregard the inconsistencies and apply the data analysis procedures available. Another is to disregard the need for systematic data analysis procedures and assume that individual

teachers will always "do the thing right." Neither of these routes is appropriate in our view.

A path we view as most appropriate is to further our understanding of classroom assessment procedures and their underlying philosophies. Then, working from the principles articulated above, modify available analysis and interpretation procedures, and develop procedures as required to build a solid analytic base for classroom assessment practice.

References

Anderson, J.O. (1990). Editorial: Assessing classroom achievement. *Alberta Journal of Educational Research*, **36** , 1-3.

Bachor, D. (1978). What does testing offer the teacher? *Special Education in Canada*, **53** (1), 18-21.

Bachor, D. (1979a). Using work samples as diagnostic information. *Learning Disabilities Quarterly*, **2** , 45-52.

Bachor, D. (1979b). Suggestions for modifications in testing low-achieving adolescents. *Journal of Special Education*, **13** , 443-452.

Bachor, D. (1990). Towards improving assessment of students with special needs: Expanding the data base to include classroom performance. *The Alberta Journal of Educational Research*, **36** , 65-77.

British Columbia Ministry of Education (1989). The Primary Program. Victoria: British Columbia Ministry of Education

Colburn, M. & McLeod, J. (1983). Computer guided educational diagnosis: A prototype expert system. *Journal of Special Education Technology*, **6** (1), 30-39.

Deno, S. L. (1985). Curriculum-based assessment: The emerging alternative. *Exceptional Children*, **52** , 219-232.

Ebel, R. (1978, March). *The case for norm-referenced measurement*. Paper presented at the American Educational Research Association Conference, Toronto, Ontario.

Educational Testing Service (1987). *ETS standards for quality and fairness*. Princeton: Educational Testing Service.

Feuerstein, R., Rand, Y., & Hoffman, M. B. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device, instruments, and techniques*. Baltimore: University Park Press.

Fuchs, L. S., Deno, S., & Mirkin, P. (1984). The effects of frequent curriculum-based measurement and evaluation on pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal*, **21** , 449-460.

McLean, L.D. (1990). Time to replace the classroom test with authentic measurement. *Alberta Journal of Educational Research*, **36** , 78-84.

Popham, W. J. (1978, March). *The case for criterion-referenced measurement.* Paper presented at the American Educational Research Association Conference, Toronto, Ontario.

Stiggins, R.J., Conklin, N.F., & Bridgeford, N.J. (1986). Classroom assessment: A key to effective education. *Educational Measurement: Issues and Practices,* 5 (2), 5-17.

Sullivan, B.M. (1988). *The report of the Royal Commission on Education, 1988: A legacy for learners.* Victoria, B.C.: Ministry of Education.

Classroom Assessment: What Research Do Practitioners Need?

Iris McIntyre
I F M Research and Evaluation Inc.

Abstract

The purpose of this paper was to present the views of field-based educators in response to the question: What research do practitioners need? Comments were gathered from a number of teachers and district administrators in British Columbia through a series of informal interviews. The results were summarized within three categories: (1) need for better links between researchers and practitioners; (2) need to build on positive attitudes towards classroom assessment; (3) need to develop new evaluation techniques to match the major curriculum changes anticipated in B.C. Although the focus was on classroom assessment, discussion included issues related to curriculum and instructional practices.

Classroom Assessment:

What research do practitioners need? The apparent disparity between research on educational measurement and classroom assessment activities has received attention in the literature (Anderson, 1989; Bateson, 1990). The national study by McLean (1985) concluded that teachers acquired their skills from other practitioners much as a craft is learned through experience. Recent research by Wilson, Rees & Connock (1989) found that teachers beyond the primary levels use evaluation to generate marks for grading purposes but seldom to monitor student progress or their own instructional strategies.

There is clearly a need to provide opportunities for teachers to gain knowledge about student evaluation (Stiggins, 1988). This is important not only to improve the quality of classroom assessment but also to ensure appropriate interpretation of test results. In light of external monitoring of student progress at provincial and district levels, teachers should be better informed about the way tests are developed and used (Rogers,1990). Because of the importance given to marks and grades, teachers need to have confidence in their own judgements about student performance and be able to communicate clearly with students and parents (Gorman, 1989).

At the same time, there has been a call for improved assessment techniques beyond the item formats traditionally used for large-scale testing (McLean, 1990; Stiggins, 1990). Work done in the United Kingdom (Black & Dockrell, 1980) and elsewhere describes new approaches to evaluation.

In order to find out the views of field-based educators on such issues, the question: What research do practitioners need? was posed to a number of teachers and people who work with teachers in the British Columbia school

system. These people were selected because of their direct involvement with provincial or district committees and who were therefore able to give opinions based on experience with groups of teachers. The interviews were loosely structured in order to allow ideas to surface without prompting.

On the whole, practitioners feel very remote from research, both research on educational topics in general and research specific to assessment. Whereas research isolates particular sections of education for investigation, teaching deals with the continuous dynamics of schooling in which it is hard to think of the parts as separate from the whole. Teachers see a real need to bridge the gap between research and practice, especially in view of the sweeping changes contained in the British Columbia Ministry of Education document titled: Year 2000: A Framework for Learning (1990). They would like to believe that research could guide classroom practice, but in the absence of such knowledge they must try to make sense of things on their own. This paper will attempt to group the teachers' comments under three headings.

Need for more direct links between teachers and researchers

How do teachers find out about research? As Ornstein (1989) pointed out: "[teachers] have little motivation for reading the research, lack research knowledge and are unable to understand the data, or feel that research is not relevant to the practice of teaching" (p. 95). Most of the practitioners interviewed mentioned that they had little or no direct contact with the literature except when actually engaged in graduate work. Motivation to read basic research in primary sources is much reduced once a teacher is outside a university environment. Even when research journals are available through local teacher libraries or resource centres, these journals are not borrowed so often as periodicals containing practical "how to" articles, according to district librarians. Teachers say that they most often hear about research second-hand from a workshop presenter, or even third-hand from someone else such as a supervisor who has attended a conference. Reviews of research are less likely to be read than articles in educational magazines where research is cited or alluded to in support of the writer's personal opinion. Teachers are presented with research findings, prepackaged and prefaced: "Research says..." without opportunity to check the statements for accuracy.

A basic need of practitioners is to have more frequent links with people who could help them sort out the salient points from all the information they are bombarded with at times. They need a kind of consumer report service to provide them with objective evaluations of current educational issues, clarify the relevance of certain research findings to their own situations by commenting on the quality of the research methodology, and point out the limitations of the studies.

The most direct links occur when teachers have an opportunity to work directly with researchers, though these opportunities are all too few. However, some teachers would like to be involved in replication of studies done elsewhere to see how the results hold up in a local setting.

Need to build on more positive attitudes toward classroom assessment

Although teachers view educational research as something remote from their classroom situations, they do hold researchers in regard as engaging in independent, autonomous, and a-political activities. By contrast, for some teachers, the word assessment is laden with connotations, some of which are not acceptable to them.

The most positive attitudes toward externally-developed achievement tests are found amongst the growing cadre of teachers in B.C. who have had an opportunity to be involved in test-development activities themselves through serving on one or other of the various provincial or district committees in connection with the B.C. Learning Assessment Program since its inception in the mid-seventies (Mussio & Greer, 1980).

A recent example occurred during the 1989-90 school year the Student Assessment Branch facilitated a process for the exchange of locally-developed assessment materials. In one case, four districts had already set up a cooperative item-development project of their own and were seeking Ministry assistance to continue their work. A group of Mathematics coordinators and their secondary teachers had started to develop a pool of math items some of which had been used in schools as unit tests or year-end tests. Other items had been written by groups of teachers meeting from time to time for after-school in-service sessions and occasionally with release time provided by the districts. The coordinators had some experience with item-writing through their own involvement on various Ministry math projects such as provincial exams, classroom achievement tests, and provincial math assessments (Taylor & Robitaille, 1987).

This is an example of a project initiated by the users themselves which in turn contributed toward a Ministry-initiative, the development a provincial item-bank (Carbol,1987). The project gathered momentum due to the enthusiasm generated by the math coordinators and the teachers.

Need to plan assessment to match new programs

The new direction in B.C. is toward ways to measure the student's individual level of attainment on specific learning objectives. The math project described above grew out of the traditional experience of measuring individual achievement against group performance. Given the nature of the subject matter in mathematics and the explicitness of the curriculum guide, it is likely that the test items can still be used by teachers to describe individual performance in terms of concepts and skills mastered. For other subject areas, and indeed in anticipated interdisciplinary approaches to learning, the development of assessment techniques will be more challenging.

How will teachers be able to evaluate their students in the new context for learning? How will they be able to meet the requirements of the new approaches to curriculum while at the same time dealing effectively with pressures from parents to report student progress? When parents ask: How well is my child doing? they usually mean: How well is my child doing in relation to all the other students? Even with the best information campaigns that the Ministry, districts

and schools can design, it will be a long time before parents fully understand and accept new ways of reporting student performance. There will be enormous pressure on individual teachers to prove that their judgements are fair and just.

In addition, teachers themselves will have to readjust to new approaches to the teaching/learning situation. Ironically, in recent years, many districts have spent a good deal of their professional development and staff development budgets introducing teachers to classroom management and instructional techniques that assume a teacher-centered and teacher-directed classroom. Most of the existing teacher-made and teacher-selected test materials are based on the assumption that a whole class (or at least a group of students) has learned a particular unit of work.

In the new situation teachers will need to be sensitive to the stages in the learning process and be able to recognize indicators of progress. The path along which an individual student's learning progresses may be continuous and sequential, or at times discursive or even regressive. In turn, this calls for curricula that are set forth with sufficient specificity to allow teachers to tell whether the student is at least heading in the right direction. No longer can the teacher be assumed to be in possession of the "right" answer as in the teacher-directed classroom.

With the introduction of the new Primary and Intermediate programs in British Columbia with their interdisciplinary approach to learning, it is likely that the language arts: reading, writing, listening and speaking, will receive emphasis in all areas of the curriculum. With this in mind, some teachers are looking toward evaluation in the language arts for clues to help with new directions in classroom assessment.

So long as measurement practice leant mainly toward multiple choice and other readily scorable item formats, language arts and English teachers were skeptical about the usefulness of such tests. But many teachers have been introduced to new approaches to teaching and evaluating student writing through curriculum implemented during the 1980s, and reinforced through the very successful Young Writers' Program. During the regular provincial assessments of written expression teachers have been involved in developing and scoring student writing using both holistic and analytic marking scales. With adequate training and monitoring, teams of teachers at the district level can also produce reliable results using such scales (McIntyre, 1987).

Crucial to this process was the development of valid criteria, specific to each writing topic. Although markers were often in agreement when asked to rank order papers, they differed about the reasons for assigning scores. The process of developing criteria was, in fact, a valuable in-service experience during which the teachers gained new insights about teaching and even modified their long-held personal beliefs.

The introduction of the writing process as a method of teaching writing meant that the teacher's role in the classroom changed from the-one-who-marks-the-papers to someone who advised, coached, modeled, encouraged, commented, and gave editorial advice. Teachers who had served on district or

provincial marking teams said that the experience gave them greater assurance when they returned to their classrooms. With the new Primary and Intermediate programs there will be many situations when the teacher cannot fall back on the"right" answer but will have to demonstrate fairness and consistency. With a generation of parents brought up on accountability, it is not likely that we shall see a return to a permissive period such as the late sixties when teachers could choose an idiosyncratic approach to evaluation. Instead, teachers will need opportunities to work together on a fairly regular basis to develop a common set of criteria in order to validate their own judgements of student work.

District administrators pointed out that policy makers have to be convinced that the cost of releasing teachers to develop evaluation materials is justified. It should be understood that such work reaches far beyond the immediate "product" —whatever form that product may take. When a group of knowledgeable, experienced, articulate teachers get together on a project they argue, discuss, criticize,explore, create. What they are doing is describing very finely what they are teaching (curriculum content); how they teach (instructional practices); and how they know what the student has learned (evaluation). During such discussions they can focus on the formative purposes of evaluation because they are not required to think about grading. Both teachers and administrators found benefit from working together on evaluation projects. It makes economic sense to encourage these opportunities which enable teachers to be more aware of the connections between curriculum, instruction and evaluation in the classroom.

Summary

The teachers and administrators interviewed all agreed that increased contact between researchers and practitioners would be beneficial, particularly in light of the new curriculum changes in B.C. They see three ways in which educational measurement could assist them: (1) providing in-service on evaluation to make teachers more knowledgeable in selecting appropriate evaluation materials and in interpreting test results; (2) assisting teacher groups in developing new evaluation materials to match the new curricula; and (3) involving teachers in research projects in local settings. In these ways teachers could become more familiar with the language used in research studies, more able to judge the relevance of results to their own situations, and more confident in the practice of teaching and evaluating their own students.

References

Anderson, J.O. (1989). Evaluation of student achievement: Teacher practices and educational measurement. Alberta Journal of Educational Research, 35(2), 123-133.

Bateson, D.J. (1990). Measurement and evaluation practices of British Columbia science teachers. Alberta Journal of Educational Research, 36(1), 45-51.

Black, H.D., and Dockrell, W.B. (1980). Diagnostic assessment in secondary schools: A teacher's handbook. Edinburgh, U.K. : Scottish Council for Research in Education.

Carbol, B.C. (1987). Issues related to the development of a provincial item banking system. Unpublished paper. Victoria, B.C.: Ministry of Education.

Gorman,. W.J. (1989). Effective Student Evaluation. Education Canada, 29(3), 4-9, 15.

McIntyre, I.F. (1987). Evaluation of a district writing program. Paper presented at the 15th Annual Conference of the Canadian Society for the Study of Education, McMaster University, Hamilton, Ontario.

McLean, L.D. (1985). The craft of student evaluation in Canada. Toronto: Canadian Education Association.

McLean, L.D. (1990). Time to replace the classroom test with authentic measurement. Alberta Journal of Educational Research, 36(1), 78-84.

Ministry of Education. (1987). Mathematics Curriculum Guide. Victoria, B.C.

Ministry of Education. (1990). Year 2000: A framework for learning. Victoria, B.C.

Mussio, J.J. & Greer, R.N. (1980). The British Columbia Assessment Program: An overview. Canadian Journal of Education, 5(4), 22-40.

Ornstein, A.C. (1989). Theoretical issues related to teaching. Education and Urban Society, 22(1), 95-104.

Rogers, W.T. (1990). Current educational climate in relation to testing. Alberta Journal of Educational Research, 36(1), 52-64.

Stiggins, R.J. (1988). Make sure your teachers understand student assessment. The Executive Educator, August 1988, 24-30.

Stiggins, R.J. (1990). Toward a relevant classroom assessment. Alberta Journal of Educational Research, 36(1), 92-97.

Taylor, A.R., & Robitaille, D.F. (1987). Classroom processes and their relationship to achievement in mathematics. Victoria, B.C.: Ministry of Education.

Wilson, R.J. (1990). Classroom processes in evaluating student achievement. Alberta Journal of Educational Research, 36(1), 4-

Emerging Needs of The Practitioner in B.C. Classrooms

Alan R. Taylor
Coquitlam School District

Background

Change has become the norm in today's society, for it seems that people scarcely have time to adjust to one innovation before the next arrives on the scene. No where is this phenomenon more evident than through the knowledge explosion currently underway. It is evident, for example, that the amount of technical and scientific data is growing at an exponential rate. At present it doubles in less than five years and it is predicted by 1996 it will double every twenty months (McKerlich, 1987).

In dealing with this deluge of information, the importance in the curriculum of data processing skills (the ability to find information, retrieve it, classify it, interpret it and report it) has become dramatically evident. Other changes, with implications for the development of new curricula, have resulted in the following shifts in focus: from teacher to learner, content to process, and passive to active learning. Each has tremendous implications for the educational system in terms of its organization and the delivery of its product.

The purpose of this paper is to discuss some of the emergent needs of the practitioner which can be met by research in dealing with these changes. In addressing this issue, it first establishes the framework within which they have evolved. Given that context the paper proceeds to identify those aspects of change which relate to curriculum and student evaluation. Second, it focuses on a number of specific questions which need to be addressed in the area of student evaluation. An approach, in which the teacher plays a meaningful role in research, is suggested as a means to address these questions at the classroom level.

Setting the scene for change

Changes in technology, the environment, and family structures and priorities, in a rapidly changing world are among the factors which gave impetus to the government and the educational community in British Columbia to reflect on the purposes and outcomes of schooling. In response, the Province of British Columbia in 1987 initiated a Royal Commission on Education. Recommendations from that Commission dealt with a wide range of issues, a number of which related directly to the curriculum and student evaluation (Sullivan, 1988). Among those with implications for researchers and with direct application at the classroom level, were the following:

> ... developmental criteria, rather than chronological age, be used in selecting the educational placement of children entering school (p.28).

... enable schools and school districts to establish ungraded primary divisions (p.28).

... a common curriculum include four categories of subject matter (p.29).

... teachers use an interdisciplinary approach in their teaching (p.31).

... provide learners ... with access to multigrade/or cross-grade classroom groupings, and assess learner progress individually (p.31).

Further to these recommendations, the report suggested that, **"The Ministry of Education should provide guidance on standards and criteria for teachers to employ in evaluating students' performance"**(p.111).

In response to the Commission's findings, the Ministry of Education developed the Year 2000 Paper (Ministry of Education, 1989) to articulate plans for follow-up activities. It proposed a curriculum and assessment framework for the British Columbia school system based on a mandate and a description of the "educated citizen". The document, which is still in draft form at the time of writing, attempts to translate many of the intents of the Commission's findings into an operational context.

Many of the changes proposed in the paper require further development at the hands of the classroom practitioner and the researcher before effective implementation is possible. A list of these, grouped under two categories: curriculum features, and evaluation and reporting, are listed below.

1. Curriculum Features

It is proposed that the curriculum :

- be learner focussed- to be "developmentally appropropriate and sequential, allows for continuous progress, provides for self direction, and is individualized as much as possible" (p.10).
- be organized according to four strands -humanities, sciences, fine arts and practical arts
- include a common component which incorporates elements of all four strands
- emphasize intended learning outcomes rather than learning activities

2. Evaluation and Reporting

It is proposed that the evaluation and reporting of student progress include the following features:

- be criterion referenced
- include a variety of assessment methods
- include continuous progress
- focus on learner profiles

Similar directions also evolved from studies undertaken in several other jurisdictions. The State of Minnesota, for example, established an Office of Educational Leadership in 1989 to establish an agenda for youth into the 21st century (Office of Educational Leadership, 1990). Many of the elements in the

plan it developed were similar to those in British Columbia. At the school district level, comprehensive plans with similar results have been developed by a number of boards, among them the Halton Board of Education (1989) in Ontario and the Coquitlam School Board (1990) in British Columbia.

Questions to be Addressed

As teachers attempt to implement extensive change flowing out of the Year 2000 Paper, additional information is essential in a number of areas. Although some are not new, the extent of proposed change and the articulated expectations, cause them to be more pressing now than ever. These issues and questions are grouped under two categories: curriculum and instruction, and student evaluation and reporting.

1. Curriculum and Instruction

It is essential that changes proposed by the Ministry of Education in the area of curriculum and instruction be based on solid research findings. Although work has been done in some areas in which change is contemplated, there remains much more to do in order to gain a clear understanding and direction. Many of these issues were listed by Costa (1989), in his Foreword to the 1989 ASCD Yearbook on current cognitive research. He identified the following questions as most pressing:

Which of tools of inquiry are important and why?

Why are modes of inquiry and thinking important in understanding and in teaching school subjects?

How do modes of thinking intersect with the knowledge base in a subject?

What instructional processes best develop subject matter concepts in students?

How much time should be spent teaching various concepts and at what developmental ages are they best taught?

How are these concepts and modes of inquiry organized to be reinforced throughout the curriculum and the school?

How can we help preservice and inservice teachers understand the concepts, modes of inquiry, and thought processes, and how to teach them?

Which of the modes of thinking and tools of inquiry are generizable to other subjects and to daily life?

How can we measure and report growth in thinking abililities? (p.vii)

2. Student Evaluation and Reporting

As a result of the proposed shift toward the student as a learner, many questions also remain to be answered in the area of student evaluation and reporting . For example, implementation of continuous progress, the use of student profiles, and reporting on what students can do, all have significant implications for procedures and practices in this area. In an attempt to come to grips with some of these, a discussion of issues related to criterion referenced measures, methods of authentic measurement, and varieties of measurement techniques follows.

(i). Criterion Referenced Measures

Movement away from norm referenced measures and toward expected standards of performance is consistent with the concept of student focused learning. Inherent in this shift, however, is need for more effective techniques by which to set standards, for it is essential that teachers be consistent and fair in their reporting of student progress. For example, clearly defined descriptors which describe student progress along a continuum need to be developed and articulated; effective methods of ongoing teacher articulation need to be established; and procedures which ensure consistency in ratings among teachers determined. The appropriate use of normative information also needs to be incorporated into this process, otherwise teacher judgments are in danger of straying apart from those of their colleagues. Once techniques and instruments are developed, there remains a crucial need for the training of teachers and administrators in assessment practices. (Stiggins, 1990)

(ii). Methods of Authentic Measurement

In a recent article, McLean (1990) defined authentic measurement as **"performance on meaningful tasks in a recognizable context"**(p.78). He provided as an example, use of a student portfolio which can be used to provide a cumulative record of student achievement for a wide variety of activities. Other examples could include the use of interviews, step-by-step processes employed by students, and "on the job" assessments for students engaged in activities out of the classroom. There is need in this area for the development of a number of ways to measure student activities in different settings and for a variety of purposes. As these techniques are developed, several ancilliary products are needed, among them appropriate instruments for the collection of information, procedures for interpretation of that data, and methods of reporting it.

(iii). Variety of Measures

For years teachers have been encouraged to use a variety of measures when they collect information about students. Most of this information, however, has been related only to intellectual development. In the proposed changes teachers will also be expected to report to parents in meaningful ways on the progress of their children in human and social development and in career development. This expectation gives added emphasis to the need for better methods of evaluating attributes of the affective domain. In addition, more work is needed on process evaluation and critical thinking. Some of this work may be done through the refinement of observational techniques and the use of holistic measures.

From the practitioner's perspective, direction related to the preceding issues is essential. For example, in the new program it is expected that inquiry methods are taught and evaluated; teaching for thinking occurs; a variety of instructional techniques to match needs of the learner are employed; and an interdisciplinary approach is used in teaching of the curriculum.

Need for Collaborative Involvement of Teachers

The extent of proposed curriculum change is staggering. Not only does it involve change in content and higher levels of cognition, but it also has implications for classroom organization, grouping of students and the use of different teaching strategies.

To effectively implement change of this magnitude the classroom practitioner is in desperate need of encouragement and direction. Yet research seldom meets these needs. For example, Perry-Sheldon & Allain (1987) contend that since the immediacy of day-to-day work calls for quick decisions by teachers, they tend to make decisions based on their own experiences and common sense rather than on those of experts or the findings of research. Cross (1987) contends it is not the case that teachers do not or will not apply findings from research in their practices, but that they cannot. This perception is supported in findings from Carr and Kemmis (1986) who suggest that teachers tend to view research as an esoteric activity having little to do with everyday matters.

At issue is the effectiveness of the top-down model currently employed in much of research. This model may be ineffective due to the isolation of teachers in the classroom, who often work without contact with colleagues or frequent supervision. Tye and Tye (1984), for example, contend that:

.... We continue to mandate changes, even though available evidence overwhelmingly indicates that reforms imposed from the top are ineffective in bringing about desired changes in schools.

The need for the direct and meaningful involvement of teachers in research was also called for in recent work done by McDonald (1989), Gage and Berliner (1989), and Atkin (1989). In a recent article, Atkin (1989) claimed that teacher-conducted research must be seen as an important responsibility of the teaching profession. In taking this position, he states that

The progress of meaningful reform will be stalled until teachers emerge from their marginal positions in the research community and become full partners in the conception and the conduct of educational inquiry.(p.205)

What direction can research gain from these observations? To begin, it is suggested that to have an effect, researchers should be as concerned with the implementation of their findings as they are with the study itself. Cross (1987) believes that the active involvement of teachers in research related to their own pratices is essential to bring about improvement in learning. This position was supported by Tikunoff and Mergendoller (1983) who saw the involvement of teachers in conducting reseach as imperative. Based on this information it is recommended that greater attention be given to the collaborative planning and implementation of research between the researcher and the practitioner.

Summary

As reported earlier, the purpose of schooling is shifting in an attempt to meet new sets of needs and expectations. Basic skills, for example, have expanded from those only related to numeracy and basic communication to include critical thinking, decision making, flexibility, and understanding and tolerance of

others. In response to these changes the developers of curriculum have redesigned their product. The curriculum has become more process oriented than content specific; it focuses more on thinking skills than subject matter; it is more selective in terms of the subject matter contained in it; and it emphasizes conceptual development and process-oriented activities.

Student evaluation is an integral component of the educational process and must be fair, consistent and supportive. As such it must be responsive to related changes in the curriculum. It is important also that student self-esteem be protected through the inclusion of tasks that students can be successful at. Gorman (1989) emphasized the importance of this feature in saying,

Evaluation is a two-edged sword that can enhance and be constructive to student learning and personality development - or it can be destructive to student learning. The choice is ours. (p.15)

Shifts in the curriculum, however, have left a void through the absence of appropriate teaching strategies, and evaluation techniques and instruments with which to measure student progress for many of the intended outcomes. This paper identified a number of needs in these areas faced by the classroom teacher in attempting to implement change. The role of research in this matrix is the development and validation of appropriate procedures, techniques and instruments. Among them, more effective means by which to measure higher order thinking skills and attributes of the affective domain; development of appropriate criterion methods to assist teachers in determining and articulating common expectations; and more effective techniques in process oriented evaluation.

References

Atkin, J. (1989). Can educational research keep pace with education reform? Phi Delta Kappan, 71, 3, 200-205.

Carr, W. & Kemmis, S. (1986). Becoming Critical. London: Falmer.

Coquitlam School District (1990). For Our Children: The Final Report of the Challenge Ahead Project. Coquitlam: Coquitlam School Board.

Costa, A. (1989). Foreword. In Resnick L. & Klopfer, L. (Eds.). Toward the Thinking Curriculum: Current Cognitive Research. 1989 Yearbook of the Association for Supervision and Curriculum Development. Alexandria: ASCD.

Cross, K. (1987). The adventures of education in wonderland: implementing educational reform. Phi Delta Kappan, 69, 496-502.

Gage, N. & Berliner, D. (1989). Nurturing the critical, practical, and artistic thinking of teachers. Phi Delta Kappan, 71, 3, 212-214.

Gorman, W. (1989). Effective student evaluation. Education Canada, Fall, 5-15.

Halton Board of Education (1989). **Toward 2000 Learning for the Future.** Burlington: Halton Board of Education.

Mc Donald, J. (1989). When outsiders try to change schools from the inside. **Phi Delta Kappan, 71, 3, 206-212.**

McLean, L. (1990). Time to replace the classroom test with authentic measurement. **The Alberta Journal of Educational Research, XXXVI,1, 78-84.**

Ministry of Education. (1989). **Year 2000: A Curriculum and Assessment Framework for the Future (Draft).** Victoria: Province of British Columbia.

Office of Educational Leadership. (1990). **State Plan Office of Educational Leadership R & D Program. (Final Draft).** St. Paul: Minnesota Department of Education.

Perry-Sheldon, B. & Allain, V. (1987). **Using Education Research in the Classroom: Fastback 260.** Bloomington: Phi Dela Kappa Educational Foundation.

Stiggins, R. (1990). Toward a relevant classroom assessment research agenda. **The Alberta Journal of Educational Research, XXXVI, 1, 92-97.**

Sullivan, B. (1988). **Royal Commission on Education. A Legacy for Learners: Summary of Findings.** Victoria: Queen's Printer.

Tikunoff, W. & Mergendoller, J. (1983). Inquiry as a means to professional growth: the teacher as researcher. In Griffen, G. (Ed.). **Staff Development: 1982 Yearbook of the National Society for the Study of Education, Part 11.** Chicago: University of Chicago Press.

Traub, R. (1990). Assessment in the classroom: What is the role of research? **The Alberta Journal of Educational Research, XXXVI, 1, 85-91.**

Tye, K. & Tye, B. (1984). Teacher isolation and school reform. **Phi Delta Kappan, 65, January, 319-22.**

Grounded Authentic Assessment and Teacher Education

Thomas O. Maguire
University of Alberta

Introduction

During the past two decades there has been a substantial increase in the amount of externally imposed assessment of student learning. Most provincial departments of education have instituted testing programs using various combinations of formative and summative tests to monitor achievement, to provide indicators of educational quality and to provide end of school certification. Recently, the Council of Ministers of Education has called for tenders to create a cross- country assessment of literacy. Beyond the provincial and national exams, some regions have participated in the periodic administration of international assessments. As we move into the 1990's a resurgence of top-down testing with its concommitent comparisons of students, teachers, jurisdictions, provinces and countries is refocussing our educational efforts. In this paper I will remind you of some of the negative consequences of this, and suggest some strategies for renewal.

Among the several research strands in the contemporary literature on achievement measurement that relate to both the consequences of provincial assessment policies, and to the suggestions for change, two were featured in last year's Victoria conference on classroom testing. The "grass roots" researcher represented by Wilson's (1990) paper, seeks to understand assessment through investigations of current teacher practice. In the next section of the paper, I will use these investigations to show how provincial activities influence instruction. The "improvement of practice" research, focusses on the validity issues of achievement assessment. One form is discussed in McLean's (1990) paper as, "authentic assessment.". I shall use this line of research in the second section of the paper to explain what assessment ought to be like. In the final section of the paper, I'll talk about implications for teacher education.

The First Part - Influences of Assessmen: Practices on Classrooms

In his study of teachers in Ontario and British Columbia, Wilson (1990) found that his modal teacher used completion, short answer or essay questions rather than multiple choice questions, they borrowed or adapted from other sources if they were in elementary schools, or built their own instruments if they were in secondary schools, and they carried out assessments for three main interrelated reasons: to determine marks, to check student progress, and to have students practice what they had learned. Near the end of the paper Wilson links teacher practise to government activity: "What actually happens in our sample of teachers is a tendency to imitate the provincial government in virtually every instrument." (p. 16) And later, "The evaluation of student achievement by teachers in classrooms occurs within a policy and procedural framework that is largely determined by outsiders" (p16). Wilson's findings support the widely held belief that provincial testing programs have a profound influence on

teacher practice.

A quick snapshot from the ATA Magazine (a publication of the Alberta Teachers' Association) will give a flavour of the Alberta context. In the March/April issue, Nemeth and Samiroden (1990) gave nine problems (Figure 1) with standardized provincial examinations. These were extracted from interviews with teachers, classroom observations, and from the opinions expressed by teachers and academics familiar with the Alberta scene. In their conclusion, the authors state,"By no means should the results of such testing be publicized or used as indicators of school or teacher performances. " (Interestingly enough the paper was written prior to the Edmonton Public School Board's decision to publish test results for each school.) Nemeth and Samiroden's thesis is that the provincial assessments and in particular the grade 12 diploma examinations represent a political approach to improving the quality of education, and what is really needed is an educational approach. They acknowledge (implicitly) that the assessment expertise of many classroom teachers is not high but that it could be improved, and in this improvement lies much of the solution to the validity/authenticity problem.

1. Diploma exams contradict the belief in continuous formative evaluation.
2. For a centralized testing program to work, tests must be normed. Norming assumes similar classroom experiences that span, in this case the province of Alberta.
3. Because there is pressure for teachers to teach them, Diploma Exams guide the curriculum and determine what is important at the expense of other goals.
4. Diploma Exams substitute content for context.
5. Test blueprints that are prepared for Diploma Exams have content area and cognitive level as their main dimensions. This classification is determined by the examiner, but the kind of thinking required is actually determined by the student.
6. Diploma Exams test for a very limited portion of the curriculum. They are geared to measure those pieces of knowledge most easily measured.
7. Diploma Exams do not correlate from year to year because the questions are revised, some new new questions are added, and the students from year to year differ.
8. Comparisons between individuals and jurisdictions are impossible. Two students who receive the same score do not necessarily understand the materials equally. Often they do not answer all of the same questions.
9. Test results only correlate with future results on similar tests; they are not accurate predictors of future success.

Figure 1. Problems associated with diploma exams (From Nemeth and Samiroden, 1990)

For these reasons Diploma Exam results are not good indicators of the quality of education.

In British Columbia the recently completed study on the impact of provincial examinations on education (Anderson, Muir, Bateson, Blackmore and Rogers, 1990) collected opinions from teachers and students from across the province. Although the analysis was cursory, the trends seemed to indicate that externally imposed exams are viewed as constraining teaching and learning practice, in terms of what is taught, how it is taught and how it is assessed. In short, the B.C. data are generally consistent with the less formal Alberta probe described above.

What are we to conclude from this information?

1. Provincial assessment programs and in particular summative programs have an important influence on instructional activities including classroom assessment practice and these directions are not always beneficial to students.
2. The assessment skills of many teachers are restricted at least insofar as they demonstrate use of a variety of techniques.
3. Despite observation 2, there is an intuitive base possessed by teachers upon which better assessment practice could be built .
4. While educational assessments at all levels seem to be purposeful, they are not obviously driven by any theory beyond the important appeal to social concern (for example, accountability).

The Middle Part - Grounded Authentic Assessment

In this section of the paper I would like to briefly reexamine and reinforce the school of thought that makes validity the fundamental concern of all achievement assessment activity. Much has been written about how the construct validity of achievement tests can be improved. A lot of this material is written from a cognitive psychology perspective (see Snow and Lohman's excellent review in Linn, 1989), but there are also important directives in the curriculum literature. Haertel, (1985), refers to these as competence theories, and although he classifies them as psychological I would call them curricular because they, "...always involve a detailed analysis of specific curricular content (p.28)."

In his paper, McLean (1990) reminds us that to be "authentic," assessment instruments must reflect, "the knowledge and the processes that form the basis of the subject matter," and the, "conditions under which the achievement normally takes place." (p80). Although not mentioned by Mclean, the entry of cognitive psychologists into the achievement assessment field (for example, Embretson, 1985) can be seen as an important step toward authentic assessment. It is a breath of theoretical freshness in a field dominated by the mathematical mustiness of item response theory.

Because authentic measurement must reflect the context of instruction, McLean warns us that it is "...impossible to produce measuring instruments that can be used in the same way across a country, or even a province(p.81)." This warning flies in the face of the political forces that seek to implement and extend

comparative testing programs at the intraprovincial, interprovincial, or international levels.

In a recent paper, Nancy Cole (1990), points out that there are philosophical issues that underlie the valid assessment of achievement. She refers to Broudy's (1988), *The Uses of Schooling* as an example of how philosophers of education have important messages for people in our field. Prompted by Cole, I read Broudy's monograph and became convinced that authentic assessment is not sufficient if we are going to restrict the term to "faithful reflection" of the knowledge and processes of the subject matter. It seems to me that *valid assessment* is authentic assessment grounded in a philosophy of education and in the psychology of development, learning and instruction. This view is consistent with Messick's (1989) extension of construct validity to include the consequential basis for interpretation.

In outlining his argument for general education, Broudy (1988) lists four uses of schooling: replicative, applicative, associative and interpretive. The replicative use of knowledge refers to the use of facts and principles just as they have been learned. In our parlance we would tie this to Bloom's (1957) knowledge objectives. Broudy points out that there is much that we learn in school that is soon forgotten in the form in which it was learned. Replicative use of learning depends to some extent on the individual. For example, I no longer remember much of the Russian vocabulary that I learned in order to satisfied the requirements for the Ph.D. On the other hand I have used the multiplication table so much that it has become overlearned.

The applicative use of schooling refers to the direct application of facts or principles to new situations. Much of technical training and apprenticeships, and some of professional education are directed toward the applicative uses of schooling. Attempts are made to show students how to structure, categorize or recognize situations where facts and principles can be applied as learned.

The other two uses, associative and interpretive, refer to how we use the associations and connotations that are built up around concepts and principles, and how we make translations and abstractions in order to live our lives. Broudy points out that the real value of education lies in the latter two uses of schooling and much less on the former. However he justifies the former by showing how they contribute to what he calls the "allusionary base," the conceptual storehouse of implicit knowledge that each of us possesses. As he notes (p. 21), "The associative resources provided by schooling and experience plus the interpretive repertoire of concepts and images consititute the allusionary base. The resources of the allusionary base are not used by simple replication of this or that school learning, and their adequacy cannot be judged by tests of replication or application. On the contrary, the success of general education is to be measured by the depth and quality of the allusionary base."

The point of describing Broudy's ideas is that if you accept them as a basis for general education, there are direct implications for the practices of assessment, both at the classroom level and externally. The philosophy should drive the instruction and assessment, and conversely the choice of assessment instruments should give clues as to the philosophy in operation. The complaint

of teachers and the observations of classroom researchers is that our tests largely neglect the associative and interpretive uses of schooling. They (the tests) paint a picture of schools that are dedicated to the replicative and applicative uses of education. Our undergraduate measurement texts are at best loosely tied to a 1956 view of educational objectives, (Bloom et al, 1956) and at worst based on the sparse logically positive dictates of the behavioral objectives movement.

The research agenda of the cognitive psychologists and others is to extend the range of assessments. But what is the basis of this extension? There is a philosophy that is implicit in this trend, and I believe that it must be explicated. I remind you that assessment, even authentic assessment, is not value neutral. Different philosophical starting points yield very different authentic measures. We must encourage assessment constructors at all levels to make their foundational beliefs and assumptions public. Authentic assessments that are derived from such a network could then be described as being grounded.

Consistent with the need to ground authentic assessment in a philosophical framework, is the importance of designing assessments that serve instruction, and using the assessment process to promote useful and responsible self-evaluation.

In a recent review of the research on the impact of classroom evaluation practices on students, Crooks (1988), reminds us that the quality of student learning is greatly influenced by the nature and frequency of our assessments. The thrust of his argument is that assessment must contribute to and not be apart from instruction. In our terms, grounded authentic assessment should encourage deep learning, it should provide effective feedback, it should promote appropriate standards, it should encourage independent learning, and it should reflect the important goals of schooling. In short, Crooks tells us to seek instructional validity in our classroom assessments. Clearly, there is a coherence among the voices of Broudy, Crooks, Cole and McLean.

To summarize the middle section of this paper, grounded authentic classroom assessment must be surrounded by and derived from an approach to education that is based upon a philosophy that is supported directly by the educators and indirectly by the socio-political views of most of the nation's citizens (social fidelity), it must be authentic, it must support instruction, and it must encourage individuals to take responsibility for their own learning.

It is clear from part one, that provincial assessments have been weighed by many in the educational establishment and found wanting. I am not convinced that the picture within classes is much better. What can teacher education do about it?

The Last Part - Implications for Teacher Education and Teacher Educators

I believe that there is a widespread feeling of powerlessness among teachers when it comes to dealing either with measurement issues in general or with externally imposed assessments in particular. This is in spite of the fact that I have never heard of a Canadian teacher being fired for the test performance of his or her students, and in spite of the fact that provincial exams are usually

created by representative teachers. If this feeling of powerlessness is as pervasive as I think it is, then we need to do something to empower teachers beginning within our teacher education programs.

In many institutions, education students are not required to take a course in classroom assessment so it is not surprising that they know very little about concepts like grade equivalence, validity and error of measurement. With a proper knowledge base, some well-honed analytical skills, and a little confidence, teachers could lead the debate on assessment and standards, and not find themselves reacting to the impositions of others.

We should begin by helping our education students to develop provisional operational philosophies of education. Arda Cole (1990) provides a good illustration of how the personal philosophies of four beginning teachers developed during their first year of teaching. It seems to me that their collection and use of assessment information should be based on this personal philosophy. Teachers with well explicated foundation can respond effectively to external information, they can work with parents, and they can make their assessments more meaningful to students and to instruction. At present, where measurement courses exist, they seem to be steeped in the technology of educational measurement. While this is important, I do not believe that it is the best starting point.

We need to teach students grounded authentic measurement strategies. One example of such a strategy that is consistent with Broudy's philosophy of general education (although it is not based on it) is Biggs and Collis's (1982). taxonomy for categorizing student responses to various educational tasks called the Structure of Learning Outcomes (SOLO). Under this system, learning outcomes are assessed by placing them into one of five levels: prestructural, unistructural, multistructural, relational and extended abstract.

The categories are defined in terms of the amount of memory capacity required in order to make the response, the kind of relating operations that are required to produce the response, and the degree of consistency and quality of closure displayed in the response. *Preoperational* responses require little memory. They may be a denial of the question or simply repeat the information contained in it. Often the responses reflect a lack of engagement between the student and the task. *Unistructural* responses concentrate on one relevant aspect of the task or concentrate exclusively on one part of the solution. Closure is premature. Students making *multistructural* responses treat aspects of the task independently, and although the memory demands are higher than at lower levels, no attempt is made to integrate the dimensions of the problem. The result is a response that may contain inconsisitencies or in which inconsistent information is ignored. Students who provide *relational* responses attend to various aspects of the task in relation to each other. Integrating themes may be used to organize the result. Inconsistencies are addressed, but no attempt is made to go beyond the boundaries specified by the task. At the *extended abstract* level the task is treated in a context that goes beyond the immediate information. Generalizations may be made to other situations of this type, but care is taken to describe the domain of generalizability. Conclusions may not be definite, but

may be qualified to allow for logically possible alternatives.

If, like Broudy, we hold associative and implicative uses of schooling to be important, then we would like our students to operate at the relational and extended abstract levels of the taxonomy. Moreover, as teachers we would constantly try to move our students from lower levels to higher levels of operation by showing them how they can use their allusionary bases to produce higher levels of discussion. From the student perspective, the SOLO taxonomy gives a structure for self evaluation. It provides them with a scale that they can use as metacognitive tool to raise their level of thought relative to a particular topic.

The SOLO taxonomy is not a panacea, rather it is an example of an aproach to measurement that could be applied in many circumstances from province to classroom. Provincial assessments that reported results according to level would act in a manner that is consistent with classroom instruction, and not be as foreign to it. Criticisms of external examinations that are based on the constraining features of these exams would be less supportable because students with a wider knowledge base (or deeper allusionary base if you prefer Broudy) would have a greater likelihood of performing at a higher level. So it is in the interests of both teacher and student to develop a wide experiential base. Yet knowledge of a common set of specific facts would not be highly reinforced.

An example of the use of SOLO in assessing skills in high school science is provided by Collis and Davey (1986). I have used it to evaluate variations on a program for gifted students (Maguire, 1988,1989). Biggs and Collis (1982) provide examples of use in history, geography, mathematics, English and modern languages. They show how the taxonomy builds on instructional theories and how instruction can be built from it. In short it provides a reasonable example of an approach to authentic assessment that fits within a philosophical framework and contributes to the instructional process.

Conclusion

There are competing demands for achievement information. We must begin to develop trust among all levels of the system so that our practices are consistent and lead to appropriate student growth and development. Obtaining an effective evaluative thrust begins with making agendas public. If achievement testing is to be used to change teaching practice, then let us say so; if it is to be used as a basis for funding, then this too should be made clear; if we want to encourage the equitable distribution of scholarships through our diploma exams, then this should be made clear. At the present time, I believe that there is much suspicion about the purposes of external exams. If we can make purposes public, then we can begin to direct developmental efforts towards agreed upon goals. Trying to make single assessments serve many masters (some of whom appear after the fact) is simply not working.

A final word must be said about "accountability." There seems to be a belief that external examinations are an instrument of accountability. The term itself is now so over used that it has lost much of its specific meaning. Moreover,

there are claims that the public demands accountablity in education, and that these demands can be satisfied by external examinations. I believe that this is largely overstated. The results of the Edmonton Public School Board's decision to publish school results was indifference. It generated two letters to the editor. There are few people who are interested in the "accountability as achievement" phenomenon but their apparent success has been due to their ability to make it seem as though they represent a grass roots movement. We have uncritically accepted studies of American schooling as being applicable here. I see no evidence of a great popular concern. If they are not asked, the members of the public seldom raise the issue. Our participation in international studies provides interesting fodder for the inner pages of the newspapers, but once we know that we stand above Burkina Faso and below Japan, a good tire fire can knock us off the page. Most of the pressure for external assessments in Canada comes from within the educational establishment. We can change this. If we can strengthen our teaching population by providing them with better assessment strategies, and we devote more of our time to research and development of grounded authentic assessment instruments, then we will end up with a more facilitating system of assessment. I think that we have the ability to do it. Do we have the will?

References

Anderson, J., Muir, W., Bateson, D., and Rogers, W.T. (1990) The impact of provincial examinations on education in British Columbia: General Report. Victoria: British Columbia Ministry of Education.

Biggs, J.B. and Collis, K.F. (1982) *Evaluating the quality of learning:The SOLO Taxonomy*. New York: Academic Press.

Bloom, B.S. (Editor) (1957) *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive Domain*. New York: McKay.

Broudy, H. S. (1988) *The uses of schooling*. New York: Routledge.

Cole, A. (1990) Personal theories of teaching: Development in the formative years. *Alberta Journal of Educational Research*. 36,(3).

Cole, N. (1990) Conceptions of educational achievement. *Educational Researcher, 19*, (3), 2 - 7.

Collis, K. F. and Davey, H. A. (1986). A technique for evaluating skills in high school, science. *Journal of Research in Science Teaching*. 23, 651 - 663.

Crooks, T. J. (1988) The impact of evaluation practices on students. *Review of Educational Research, 58*, 438 - 481.

Embretson, S. (Ed.) (1985) *Test design: Developments in psychology and psychometrics*. New York: Academic Press.

Haertel, E. (1985) Construct validity and criterion referenced testing. *Review of Educational Research, 55*, 23 - 46.

Linn, R. (1989) *Educational measurement (3rd Edition)*. New York: MacMillan.

Maguire, T. O. (1989) Gifted education in the Calgary Public Schools. Calgary: The Calgary Board of Education.

McLean, L. D. (1990) Time to replace the classroom test with authentic measurement. *Alberta Journal of Educational Research*. 36, 4 - 17.

Messick, S. (1989) Validity. Chapter 1 in Linn, R. L. *Educational measurement (3rd Edition)*. New York: MacMillan.

Nemeth, J. and Samiroden, W. (1990) Diploma exams: A political approach to evaluation. *The ATA Magazine, 70*, (3), 28 - 31.

Snow, R. E. and Lohman, D. F. (1989) Implications of cognitive psychology for educational measurement. Chapter 7 in Linn, R. L. *Educational measurement (3rd Edition)*. New York: MacMillan.

Wilson, R. J. (1990) Classroom processes in evaluating student achievement. *Alberta Journal of Educational Research*, 36, 4 - 17.

What Skills Do Teachers Need in Educational Testing?

Ronald K. Hambleton
University of Massachusetts at Amherst

Educational testing in the 1990s is going to look very different from the testing of the last couple of decades. In the 1990s, more educational tests are going to be performance-based and will be more likely to measure higher-order cognitive skills than their predecessors. Multiple-choice testing is not likely to be discontinued in American or Canadian schools (nor should it be), but this test format is likely to be balanced in school, state or provincial, and national testing programs by more direct measures of assessment such as writings tasks, performance tests, projects, and portfolios.

Our predictions will hardly come as a surprise to teachers and administrators. The educational journals have been filled recently with debates about the merits of current testing programs; many school districts and provincial/state departments of education have been reshaping their testing programs (e.g. Valencia, *et al.*, 1989; Roeber & Dutcher, 1989); and the major test publishers, themselves, appear to be following the trends by including more advanced skills in their standardized achievement tests. Import U.S. national testing programs, such as the National Assessment of Educational Progress (NAEP)., anticipate more use of open-ended item formats and the assessment of higher-order skills in the 1990s to fall in line with new curriculum specifications (Collis & Romberg, 1991; National Council of Teachers of Mathematics, 1989). The directions for testing in the 1990s seem clear.

Where is the impetus for change in testing practices coming from? Undoubtedly, the impetus for changes in testing as well as in school organization, curricula, teacher training, and so forth, is coming from the widely held view that schools are not doing the job they should be doing. Consider some of the recent headlines from U.S. newspapers: "Educators say public schools need overhaul, not reform," "U.S. youth fail math test," and "American schools perpetuate failure." Or consider the findings in a recent international study of eight grade mathematics and science skills where American students placed last (Lapointe, Mead, & Phillips, 1989).

The six national educations goals prepared by President Bush and the governors in 1989 are a response to the problems that have been identified in the United States and are intended to improve education substantially by the year 2000. The goals are (a) to prepare preschoolers for learning by improving their health care and nutrition, (b) to increase the high school graduate rate, (c) to make Americans the best in the world in science and mathematics, (d) to reduce the adult rate of illiteracy to zero (from the current level or 13%), (e) to make every school free of drugs and violence, and (f) to require students in grades 4, 8, and 12 to demonstrate competency of higher-level cognitive skills in history, math, science, geography, and English. These goals are ambitious, and, to achieve them (especially goals c and f), one of the main recommended activities

is to assess students not by their ability to memorize facts and details but by their reasoning and problem-solving skills. As a result, education testing in the United States will need to change: change what is measured (to reflect curriculum changes) and use an expanded number of item formats to enhance the validity of educational assessments. Similar changes can be seen in the testing programs in many Canadian provinces.

The new era in testing has been labelled authentic assessment. Though we dislike the term itself because we believe it denigrates 80 years of important advances in psychometric research (see, for example, Linn, 1989; Hambleton & Zaal, 1991), many of the ideas underlying the new testing movement appear to be sound, and the results will be better educational assessments for policy makers, school administrators, teachers, students, and their parents. Objective testing, as represented by the multiple choice and true-false formats can carry a significant part of the assessment load (see, for example, Farr, Pritchard, & Smitten, 1990), but certainly not all of it. As more and more use is being made of test results in educational accountability (e.g., evaluating teachers, schools, programs, and even states - see, for example, the 1990 NAEP Trial State Assessment at Grade Eight), those affected by the results will want to be sure that the tests themselves are fair and accurately measure what students are learning. Even the most ardent supporters of standardized achievement tests would not claim that these tests measure more than a fraction of what schools expect students to learn or that the tests can measure, with the multiple-choice format, all of the important higher-order cognitive outcomes.

If teachers are going to teach to the skills covered by tests (and there is substantial evidence that they do already), current thinking is that tests should measure what is really important in a curriculum (Madaus, 1988). Then, teaching to the test is constructive and desirable. Shepard (1989) has even argued that assessments should approximate the learning tasks of interest, so that, when students practice for the test, some useful learning takes place. The same argument was made for criterion-referenced tests in the 1970s and 1980s. Load up tests with the skills that students are expected to learn. Teaching to the test, then, is equivalent to teaching the curriculum. What distinguishes the current movement for authentic testing from the earlier one for criterion-referenced testing are two features: (1) curriculum specialists are arguing for the teaching and assessment of integrative skills, e.g., whole language, that are typically broad and higher-order, rather than the narrower and lower-level discrete skills (i.e., behavioral objectives) that have been popular in the last 20 years, and (2) there is more emphasis in authentic testing on direct measures of the skills of interest. Criterion-referenced testing and authentic testing are not at odds. In fact, authentic testing is simply a "fresh face" on criterion-referenced testing which highlights the need for the assessment of higher-order cognitive skills with more direct measures of assessment.

The remainder of this paper is divided into sections: in the first, the term authentic testing is designed and the case for this type of assessment is considered. In the second part, the testing skills teachers and administrators will need to implement authentic testing successfully are considered.

Authentic Testing

We begin by noting that a single definition of authentic testing has not been agreed up by its many advocates. Fortunately, the main features of most currently popular definitions are clear (see, for example, Horvath, 1991):

1. Authentic tests are intended to assess what it is that students know *and* can do, with the emphasis on "doing".

2. Authentic tests should use direct methods of assessment (e.g., writing samples to assess writing, and oral presentations to assess speaking skills) whenever possible.

3. Authentic tests should have a high degree of realism about them. That is, in reading assessments, students would be expected to read reasonably lengthy passages (perhaps several pages) prior to answering questions, and, in mathematics tests, students would be expected to work with rulers, protractors, calculators, and so forth, in solving mathematics problems.

4. Authentic tests might involve (a) activities for which there is no correct answer. (b) assessing groups rather than individuals (e.g., a group putting on a play). (c) testing that would continue over an extended period of time, or (d) self evaluation of performances, projects, and so forth.

The first feature is *not* unique to authentic testing. In fact, it is a central feature of criterion-referenced testing (see, for example, Popham, 1978). The second feature is not unique, either, though it would be correct to say that authentic testing advocates aspire to see a great deal more use of performances assessments than has been the norm (see Linn, 1991; Linn, Baker, & Dunbar, in press). For example, in British Columbia, at the lower grade levels, many of the important objectives or school outcomes are being measured with performance assessments. Forman and information observations, qualitative analysis of student performances and products, oral questioning, and analysis of student records, are just a few of the other assessment methods that have been suggested in the measurement literature.

The third feature is very important. The goal is to make testing more like instructional activities than like highly structured tasks in which answer choices are provided. Some changes can be made to multiple-choice testing by providing more realistic stimuli such as longer, more interesting, and thought-provoking passages, and the use of more "application of knowledge" questions, but there are practical limits as to what can be done. The use of non-multiple-choice formats holds more promise for assessing higher-order thinking skills. Authentic tests might require students to prepare a research paper, conduct an experiment, participate in a debate, and so forth.

The fourth feature is the most problematic because it makes it nearly impossible to produce standardized directions, scoring, and interpresentations. Such a feature may be attainable within classroom testing practices (if teachers are fully trained in performance testing methods) but it will be difficult to achieve in school, and certainly in provincial/state and national testing programs. The feasibility of more performance-based testing at the provincial/state and national level is a hotly debates topic at the moment.

Testing Competencies for Teachers

The topic of what to teach teachers about testing has finally become a big issue in education (e.g., Schafer, 1991). The issue is big enough that the two largest teacher unions in the U.S., the National Education Association (NEA) and the American Federation of Teachers (AFT), joined with the National Council on Measurement in Education (NCME) to produce the AFT, NCME, NEA *Standards for Teacher Competence in Educational Assessment of Students* (1990). The *Standards* were prepared to service a number of purposes:

a. a guide for teacher educators who design and/or approve teacher education programs.

b. a basis for teachers conducting a self-evaluation of their educational testing skills.

c. a guide for the design of testing workshops for teachers.

d. a directive to educational measurement specialists and teacher trainers to broaden their conception of student assessment and convey this broader conception in their research, writing, and teaching.

The *Standards* are more powerful than other efforts to develop testing guidelines (e.g., Popham & Hambleton, 1990) because they have the full backing of the two teacher unions. Also, unlike some of the other efforts, these *Standards* were widely circulated and reviewed prior to their publication and were comprehensive in the sense that a broad definition of student assessment was adopted. Also, the *Standards* cover the complete set of teacher activities where student assessments are done. These activities are: (a) activities prior to instruction, (b) activities occurring during instruction, (c) activities occurring after the regular segment of instruction, (d) activities with a teacher's involvement in school and school district decision-making, and (e) activities associated with a teacher's involvement in a wider community of educators.

The testing competencies for teachers were organized by AFT, NCME, and NEA into seven broad areas, with each area further described by a set of skills that teachers would need to be proficient (see AFT, NCME, & NEA, 1990). The competencies are:

1. Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.

2. Teachers should be skilled in developing assessment methods appropriate for instructional decisions.

3. Teachers should be skilled in administering, scoring and interpreting the results of both externally produced and teacher-produced assessment methods.

4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, and developing curriculum and school improvement.

5. Teachers should be skilled in developing valued pupil grading procedures which use pupil assessments.

6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.

7. Teachers should be skilled in recognizing unethical, illegal, and otherwise

inappropriate assessment methods and uses of assessment information. (AFT, NCME, NAE, 1990)

By starting with a broad (and appropriate) definition of student assessment:

Assessment is defined as the process of obtaining information that is used to make educational decisions about students, to give feedback to the student about his or her progress, strengths, and weaknesses, to judge instructional effectiveness and curricular adequacy, and to inform policy (AFT, NCME, NEA, 1990, p. 1).

AFT, NCME, and NEA make the strong case for "authentic testing" without the rhetoric which is often associated with the plea. No "put-down" of norm-referenced testing, no criticism of current testing practices, and no direct challenge to the multiple-choice format or other objective item formats is offered. But the case for a broader set of item formats in school testing is clearly articulated in all of the competencies but especially in (1), (2), (3), and (4).

Of course, arguing for the competencies is only part of the solution. Training teachers to meet the competencies is likely to be difficult and time-consuming. Clearly then, substantial changes in the pre-service training of teachers are called for to meet the requirements of these new *Standards*.

Summary

There is a need in education for new educational assessments to measure important skills that cannot be measured well by traditional forms of assessment. At the same time, alternative forms of assessment, per se, will not be necessarily be useful. Their existence certainly does not equate to usefulness. The same basic concerns for test standardization, reliability, and validity are still operative. Perhaps educators would be wise to identify those testing situations where authentic assessment can strengthen their testing program; then to allocate sufficient time and resources to their development to insure that these new tests will have the psychometric characteristics to meet the expected need; and, also to allocate sufficient time and resources for test administration and test scoring to insure these essential jobs are done well. Finally, it is hoped that the testing competencies recommended by AFT, NEA, and NCME will be taken seriously so that teachers will be prepared to contribute successfully to the authentic testing movement.

References

AFT, NCME, NEA (1990). *Standards for teacher competence in educational assessment of students*. Washington, DC: NCME.

Collis, K. & Romberg, T. A. (1991). Assessment of mathematical performance: An analysis of open-ended test items. In M.C. Wittrock & E.L. Baker (Eds.). *Testing and cognition* (pp. 82-130). Englewood Cliffs, NJ: Prentice-Hall.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement*. 27, 209-226.

Hambleton, R. K., & Zaal, J. N. (Eds.). (1991). *Advances in educational and psychological testing*. Boston, MA: Kluwer Academic Publishers.

Horvath, F.C. (1991, April). *Assessment* in Alberta: Dimensions of authenticity. Paper presented at the meetings of NATD/NCME, Chicago.

Lapointe, A. E., Mead, N. A., & Phillips, G.W. (1989). *A world of differences: An international assessment of mathematics and science* (Report 19-CAEP-01). Princeton, NJ: Educational Testing Service.

Linn, R. L. (Ed.). (1989). *Educational measurement (3rd ed.)*. *New York: Macmillan*.

Linn, R. L. (1991, April). *Alternative forms of assessment*. Paper presented at the meeting of AERA, Chicago.

Madaus, G. F. (1988). The influence of testing on the curriculum. In I. N. Tanner (Ed.). *Critical issues in curriculum: 87th yearbook of the National Society for the Study of Education* (pp. 83-121). Chicago: University of Chicago Press.

National Council of Teachers of Mathematics. (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J., & Hambleton, R. K. (1990). Can you pass the test on testing? *Principal*, 38-39

Roeber, E., & Dutcher, P. (1989). Michigan's innovative assessment of reading. *Educational Leadership*, 46(7), 64-70.

Schafer, W.D. (1991). Essential assessment skills in professional education of teachers. *Educational measurement: Issues and pratice*, 10, 3-6, 12.

Shepard, L.A. (1989). Why we need better assements. *Educational Leadership*, 46(7), 4-9.

Valencia, S.W., Pearson, P.D., Peters, C.W., & Wixson, K.K. (1989). Theory and practice in statewide reading assessment: Closing the gap. *Educational Leadership*, 46(7), 57-63.

Making Assessment Training Relevant for Teachers

Richard J. Stiggins
Northwest Regional Educational Laboratory

In a study recently completed by NWREL, we found that less than half of the largest undergraduate and graduate teacher training programs in our six-state region offer the option of assessment training to their students. Further, less than a quarter of these programs *require* the successful completion of this course by their students (Stiggins and Conklin, 1989). Thus the vast majority of the teachers coming out of these programs and going to work in the classrooms of our region do so having received virtually no guidance whatever in assessing student achievement—an activity that will command as much as a third to a half of their professional time.

This is not a new problem. We have known for decades that teachers and administrators alike are inadequately trained in assessment. Yet despite research-based reminders of this fact about once every ten years for the last 50 years, nothing has changed. I submit that this does not reflect a problem of benign neglect. Rather it suggests that there are purposeful forces at work within and outside of the education community to prevent assessment training from becoming part of the professional preparation of educators.

Since I have speculated elsewhere about the nature of those forces (Stiggins, in press), I will not address all of them in detail here, except to say that I am convinced that one of the primary causes of the absence of assessment training in the teacher training curriculum has been the chronic and deep-seated mismatch between what teachers need to know about assessment and the content of assessment courses when they are offered. In other words, I submit that we in the measurement community have failed to convince those who manage and direct teacher and administrator training programs that we have anything of relevance to share with practitioners.

For this reason, my colleagues and I have reflected upon, researched, developed and field tested various strategies for enhancing the relevance of our assessment training. In the course of completing this work, we have systematically analyzed the task demands of classroom assessment, constructed a wide variety of workshops for teachers, presented those training sessions to tens of thousands of educators, experimented with various modes of presentation (including video), and followed up to determine reactions to and impact of our training. Out of these various activities, I have gleaned several keys to successful assessment training for teachers. My purpose in this paper is to share eight of those keys to success.

Classroom assessment training will be relevant and helpful for teachers when the following standards are met (in order of importance):

1. the trainer understands the realities of classroom life,
2. teachers come to understand that student academic and personal well-being hinges on the quality of the teacher's own classroom assessments,

3. the trainer understands the achievement targets teachers must assess,

4. the trainer understands the broadest possible range of available assessment methods and can align them appropriately with targets,

5. the trainer can compromise between the real and the ideal in defining and helping teachers come to terms with issues of assessment quality,

6. the training reveals to teachers the fact that systematic classroom assessment can make their job faster, easier and better, in that order,

7. the trainer relies on instructional strategies that involve teachers in developing and conducting assessments rather than hearing about them, and

8. both the trainer and the training is geared to be effective both in preservice and inservice training contexts.

Let me expand on each of these, detailing specifically how to meet each of these standards and why it is critical to do so.

1. *Know Classrooms*

The trainer must understand the realities of life in the classroom as those realities are seen and experienced by teachers. The trainer must tune into and appreciate the complex and very demanding assessment requirements faced by teachers, who may be teaching from 30 to 200 students a day, making decisions at the rate of one every two to three minutes, and are needing to filter through the constant flood of information coming from students to find those bits of information that deserve the teacher's careful attention. Further, trainers must understand the full range of uses of classroom assessment, ranging from decision making to instructional uses to classroom management uses. They must understand the interpersonal facets of classroom assessment, the opportunities offered by the assessment context where 99.9% of all school assessment takes place, and how classroom assessment differs from its more visible cousin, large-scale, standardized testing.

Trainers must possess this working knowledge of assessment life in classrooms, in order to form a clear and highly-differentiated vision of what it is about assessment that teachers need to know. But more importantly, trainers who can demonstrate sensitivity to life in classrooms can establish credibility with teachers as a knowledgeable source of sound, practical ideas about how to manage classroom assessment environments effectively. Without that credibility, the trainer's message simply will not get through.

2. *Establish the Importance*

Trainers must appeal to teachers to care about quality classroom assessment. This can be accomplished first by appealing to teachers' feelings about the well being of students, and second by involving principals and other supervisors, who evaluate the performance of teachers.

With respect to student well being, teachers respond to the argument that they are in fact at the center of the assessment process that feeds virtually all of the decisions that exert the greatest influence of learning and academic self concept. To illustrate, teachers themselves use the results of day to day assessments to diagnose student needs, group students, grade students and

evaluate instruction. Obviously, these decisions are critical to student learning.

In addition, the teacher's classroom assessments provide the information students use to make some crucial decisions for and about themselves, such as setting their own expectations of themselves, and deciding what to study, when to study, with whom to study, whether to study, indeed whether to care about and expect to be successful in school. And on top of these, parents rely on day to day classroom assessment results to inform decisions about what expectations to communicate to their child, whether and how to assist the child with homework and establish support systems at home, and how to allocate family resources for education. Taken together, decisions made by teacher, student and parent drive the teaching/learning process by determining what kids learn and how they feel about it. All of these decisions rest on the quality of the teacher's day to day assessments of student achievement. Teachers can appreciate this fact when it is pointed out to them. And when they do realize it, they become eager and willing students of sound classroom assessment methods.

However, if further motivation to care about high-quality classroom assessment is needed, it can be generated by involving principals or other supervisors in the process of promoting sound assessments. Instructional leaders need to be in a position to lead in all aspects of the instructional process, including the one that commands as much as a third to a half of a teachers available professional time: assessment. Principals are responsible for evaluating the performance of teachers and for promoting the professional development of those teachers who need to enhance their skills. If teachers see that their supervisors care about and intend to evaluate classroom assessment competence, the teachers will see the value of expanding their capabilities in this important performance arena.

The well-being of students turns on the quality of teachers' assessments of student achievement. Teachers are not currently trained to maximize that quality. They must care about this and make quality a priority if quality is to be achieved.

3. *Know the Targets*

For decades now, the measurement community has dictated the format of assessment in schools. The traditional dictum has been "This is the assessment method, educators—the multiple-choice test item. Now make your targets fit this method." Recently, however, educators have come to realize that most of the achievement targets they hold as valuable for their students simply cannot be translated into the "accepted" format. Now they have begun to say back to the measurement community: "These are our achievement targets—they are many and diverse and do not fit the traditional method. Give us methods that fit our targets." To be effective in offering assessment training to teachers, trainers must understand those targets and know how they align with assessment methods.

Many kinds of targets are valued. Educators want their charges to master substantive subject matter knowledge, demonstrate that they can use that knowledge and their higher order thinking skills to solve problems, exhibit certain kinds of achievement-related behaviors, create achievement-related

products that possess certain specified attributes, and attain certain affective targets. The successful classroom assessment trainer understands each of these. More specifically, if the achievement target is writing, reading, speaking, doing good science, doing good math, etc. the qualified assessment trainers must possess at least a working knowledge of such targets if they are to assist teachers in linking those targets to sound assessments.

Certainly, it is not reasonable to expect assessment training to address all relevant targets. Methods courses must take lead responsibility for this facet of teacher preparation. However, it is reasonable to expect assessment specialists and methods instructors to work in a close partnership to train teachers to assess all relevant targets.

4. *Know the Methods*

The assessment of student achievement in the classroom requires the application of a wide variety of measurement methods, including paper and pencil instruments, performance assessments (assessments based on observation and judgment), and personal communication with students. Further, each of these methods includes both objectively-scored and subjectively-scored alternatives, many of which come to the teacher with the textbook and some of which are developed by the teacher. In fact, teachers have literally dozens of assessment options at their disposal for assessing student attainment of valued achievement targets.

Assessment training is properly focused when the trainer understands all of these options in terms of the kinds of targets each can reflect, the strengths and limitations of each, the keys to their effective development and use, and the pitfalls to sound assessment in each case. Assessment training is made relevant, when the trainer can communicate these options to teachers effectively in practical terms, using language and illustrations teachers can understand.

Without knowledge of all available assessment tools, it is impossible for teachers to match their assessment methods to their valued targets on a day to day basis.

5. *Compromise on Quality*

High-quality assessment is critical to student well being. Accepted psychometric guidelines for treating issues of assessment quality hold that the assessor (a) understand the definitions of various kinds of validity and reliability, (b) know how to maximize the quality of their assessments (i.e. control for sources of invalidity and unreliability), and (c) know how to generate statistical estimates of quality, so as to be able to defend their assessments as meeting accepted standards. These guidelines represent the ideal when it comes to addressing quality control issues in assessment.

However, given the demands of life in classrooms, the highly-technical nature of the definitions and statistical estimation procedures mentioned above and the extremely limited amount of time available for assessment training, it simply is not realistic to hold teachers accountable for meeting all of these guidelines. Assessment training is made relevant for teachers when the trainer

is able to compromise between the real and the ideal in preparing teachers to address issues of quality.

Teachers need to know what can go wrong when they attempt to translate a particular achievement target into an assessment and they need to know how to prevent those problems from arising (entry b from the list above). They need to know about sources of measurement error that can arise from problems with the assessment exercises, problems with the student, problems with the assessment environment, and problems with the scoring process, particularly when scoring is subjective. Given this information, they can *maximize* the quality of their assessments. They do not need to know definitions of various forms of validity and reliability (entry a above) or how to estimate (entry c above) to be able to avoid the problems that reduce assessment quality. Besides they have neither the time nor the technical supported needed to deal with statistical procedures. Thus instruction on how to carry out such procedures wastes extremely valuable and limited classroom assessment training time.

Teachers need to develop and use the best quality assessments they can. To do so, they need to know the potential pitfalls and how to avoid them in very practical, specific terms that translate into efficient actions they can take to maximize quality.

6. *Reveal the Economies*

I have yet to meet a teacher who is looking for more work to do. In fact, teachers manage very full and demanding schedules. So if the well-meaning assessment trainer arrives on the scene demanding that teachers add a whole new array of activities related to better assessment to already full agendas, at best they will not be listened to and at worst they might suffer embarrassing abuse from over-worked educators. In this context, the challenge is to find some way to influence how teachers spend their professional time. One excellent way to do this is to promise teachers that systematic classroom assessment will make it possible for them to do their job faster, easier and better, in that order.

Not only must we show them that assessment in not a time eater, but we must convince them that careful assessment can be an immense time saver for them. This can only be accomplished by developing a collection of time and labor saving ideas to share during training, of which there are many: having clear targets makes assessment easier, using tables of test specifications saves paper and pencil test construction time, clear performance criteria makes communication with students and parents faster and easier, planful grading practices allows one to take advantage of efficient sampling strategies, etc. When such tactics are the highlights of training, assessment training is relevant for teachers.

7. *Teach Well*

Assessment training also is made most relevant when the trainer takes advantage of instructional strategies that engage teacher is doing assessment—that model assessment methods in practical terms for teachers to touch. These can involve assessment simulations, evaluations of previously-developed assessments, applications of the principles of sound assessment to real-world

problems, and assignments that make teachers observers of the assessment world around them—that is, make them critical consumers of the assessments that guide the lives of the students around them. These strategies engage teachers.

Further, in graduate and undergraduate assessment courses, where the trainer must also assess, evaluate and grade the achievement of teachers or perspective teachers, it is also very useful to use course projects and examinations as opportunities to model many different forms of assessment. In this way, teachers can experience both the doing of the assessment and the effects of the results on them as students. Trainers do well to debrief these experiences with their charges, drawing inferences for teachers to carry back to their own classrooms. Such sensitizing experiences make training relevant.

8. *Attend to Preservice and Inservice*

It is completely inappropriate for competent classroom assessment trainers to think of their work as involving only the improvement of university-based assessment courses. While this is a very high priority and should command a great deal of our attention and energy, we must also face the fact that the vast majority of teachers currently practicing in schools graduated from programs that offered them no assessment training. Essentially, we have an entire national faculty in need of relevant, helpful assessment training. Our limited training resources must be distributed to serve both preservice and inservice training priorities.

This relates to the issue of making training relevant for teachers because it forces the trainer out into the real world and forces them to develop professional development experiences that work in the inservice market—a tough place to succeed by any standard. Consider the challenges we face as inservice trainers: Establish our credibility as a source of useful ideas by presenting training on a topic that teachers often know little about, are very anxious about and that is risky for teachers, in brief but productive workshops presented after school, when teachers are exhausted. Further, we must make the training events work, in the sense that they actually do change teachers' assessment values and practices. Training developed to work in this context can work anywhere. So this represents an excellent place to start.

Conclusion

We live in a society that has become obsessed with the attainment of educational outcomes. Evidence of that obsession abounds. National and international assessments of outcomes that were unheard of a few years ago now are commanding hundreds of millions of dollars of our education resources. Statewide assessments have grown in number over the past decade to include nearly every state and to claim total costs that far exceed the costs of the national and international assessments. The use of standardized achievement batteries in local district wide assessment programs continues to grow to record levels.

In this kind of environment, it is both paradoxical and very troubling to realize that we also are a society that is almost completely illiterate with respect

to the assessment of the outcomes we value so much. That illiteracy includes practitioners in our classrooms, building offices, district offices, and boardrooms. If we are to reach our goal of having all students attain our valued educational outcomes, we must develop and present relevant, helpful assessment training for all, beginning with teachers.

References

Stiggins, R.J. (in press). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice.*

Stiggins, R.J. & Conklin, N.Faires (1989). Teacher Training in Assessment. An unpublished paper. Portland, OR: Northwest Regional Educational Laboratory.

A Call for Measurement Standards in Canada

W. Todd Rogers
University of Alberta

OLD SCONA IS HEADS ABOVE PUBLIC SCHOOL COUNTERPARTS
(Panzeri, April 17, 1990)

TEST SCORES MISLEADING PRINCIPALS SAY

SCHOOL RANKINGS IRK EDUCATORS

Public education is not served by ranking schools according to the results that students achieve on diploma exams, says an Edmonton high school principal.

Other principals offered similar sentiments in response to an article in Tuesday's *Journal* that ranked high schools on the basis of last year's diploma exam results. (Panzeri, April 18, 1990)

At last year's Conference on Classroom Testing held at the University of Victoria, I concluded my paper with the following paragraph:

Despite this strong support for testing and assessment in British Columbia, the testing community in this province as elsewhere (see Madaus, 1985) must continually monitor the way in which tests are developed and used. We should remain mindful of the concerns raised by those who question the "overuse" of tests. For example, we saw two of the advisory research teams to the Royal Commission disagree on the continued use of Grade 12 government examinations. This apparent contradiction is centered on the widespread concerns that the content of the examinations had unduly influenced the curriculum and that teachers teach only what is considered likely to be tested. In the same vein, tests are frequently criticized for apparently focusing only on that which is easy to test: recall of factual information, application of routine algorithmic procedures, and the like. We need to be aware of such concerns. We must be prepared to give the consumers the protection they deserve by being mindful of testing, assessment, and evaluation as practiced and to reassert the fallibility of measurement even in the face of those who are obsessed with the infallibility of numbers. (Rogers, 1990, p.63)

The focus of the present paper is upon this call for consumer protection. More specifically, what steps need to be taken to provide protection against potential misuse of testing and test results?

But before addressing this question, is there cause for alarm?

TEACHERS, PRINCIPALS CRITICIZE DECISIONS TO MAKE SCORES PUBLIC.

(Panzeri, April 23, 1990)

Illustration 1.

This headline and the ones quoted at the beginning of this paper suggest that there may indeed be cause for alarm. On April 17, 1990 *The Edmonton Journal* published the final blended marks, defined as the average of the 1989 Alberta provincial diploma examination score and the grade awarded by the school, achieved by each of 12 high schools in the Edmonton Public School system. The final marks for the seven subject areas tested (English 30 and 33, and Social Studies, Mathematics, Biology, Chemistry, and Physics 30) were accompanied by their corresponding ranks (Panzeri, April 17, 1990). Six days later the Journal published school-by-school results of the 1989 Alberta provincial achievement tests administered at Grades 3 (Language Arts), 6 (Social Studies), and 9 (Science). In this case the mean percent and number of students who wrote the examination were listed for each public school in Edmonton with students at the corresponding grade levels. Unlike the situation for high schools, school ranks were not reported (Panzeri, April 23, 1990). Other than the following statement which appeared at the top of the table in which the achievement results were reported, no other information or data were reported:

Edmonton public school-by-school results of the 1989 Alberta achievement tests. Achievement tests are not used to grade either teachers or students. They are used as a measure of whether the curriculum has been taught and learned in the areas tested. If the number of students, in parentheses, is less than 25, the results should be interpreted with caution. (Panzeri, April 23, 1990).

As published, this reporting of the school-by-school results to the public fails to meet the standards for reporting set forth in the *Standards for Educational and Psychological Testing* (American Psychological Association, 1985) and in the *Code of Fair Testing Practices in Education* (American Psychological Association, 1988), part of which is shown in Table 1. What is the evidence to support his claim?

1. Despite the statement by Alberta Education, the test developer, that the results of its Provincial Diploma Testing Program and its Achievement Testing Program are not intended to be used to make comparisons among schools (Alberta Education, 1989a, p. 107; 1989b, p.3), the test user employed the test results for this specific purpose. Concerned with trends that showed their public schools performed consistently below provincial averages, the central school administration of the Edmonton Public School system announced in early January, 1990 that they would release provincial examination scores on a school-by-school basis in an attempt to reverse this trend. Prior to this time individual schools had provided to their parents the results obtained by the school.

Table 1

Reporting Standards

A. **Standards for Educational and Psychological Testing**
(American Psychological Association, 1985)

Standard 15.10 Those responsible for testing programs should provide appopriate interpretations when test score information is released to students, parents, representative teachers, or the media. The interpretations should describe in simple language what the test covers, what scores mean, common misinterpretations of test scores, and how scores will be used. (Primary) (p.84)

B. **Code of Fair testing Practices in Education**
(American Psychological Association, 1988, p.2)

Test developers should help users interpret scores correctly	Test users should interpret scores correctly
Test Developers Should:	**Test Users Should:**
9. Provide timely and easily understood score reports that describe test performance clearly and accurately. Also explain the meaning and limitations of reported scores.	9. Obtain information about the scale used for reporting scores, the characteristics of any norms of comparison group(s), and the limitations of the scores.
10. Describe the population(s) represented by any norms or comparison group(s), the dates the data were gathered, and the process used to select the samples of test takers.	10. Interpret scores taking into account any major differences between the norms or comparison groups and the actual test takers. Also take into account any differences in test administration practices or familiarity with the specific questions in the test.
11. Warn users to avoid specific, reasonably anticipated misuses of test scores.	11. Avoid using tests for purposes not specifically recommended by the test developer unless evidence is obtained to support the intended use.
12. Provide information that will help users follow reasonable procedures for setting passing scores when it is appropriate to to use such scores with the test.	12. Explain how any passing scores were set and gather evidence to support the appropriateness of the scores.
13. Provide information that will help users gather evidence to show that the test is meeting its intended purpose(s).	13. Obtain evidence to help show that the test is meeting its intended purpose(s).

With the central staff announcement, school results would now be provided as a set. Following the initial announcement and in view of strong opposition from teachers and principals, there followed a period of intense discussion and debate involving trustees, central district staff, principals, teachers, members of the media, and interested parents and members of the public. Of concern was the permissibility and validity of such a release. Following receipt of a legal opinion that the School Act (Province of Alberta, 1988) granted authority to the Board to release school results, the Board in April, 1990 authorized the release, school-by-school, of the 1989 blended diploma marks and the 1989 achievement test scores.

2. The results as published in the *Journal* in essence do not even serve their intended purpose, which is to tell how well or poorly individual schools are performing. What, then, is missing or wrong?

First, and of paramount importance, is the questionable validity of the use of diploma marks and achievement test scores to make comparisons among schools. The tests used were not designed to yield all the data needed to make complete and valid comparisons. Schooling is more than the relatively narrow range of learning outcomes assessed by the tests considered. Although these outcomes should be taught and learned, schools are responsible for learning outcomes that extend beyond those assessed by current provincial testing programs in Alberta. Simply put, as the only source of information current examinations tell only part of the story; telling information on other prescribed learning outcomes is missing.

Second, valid interpretation of a school result requires consideration of other factors (including various demographic and opportunity-to-learn variables) which have been shown in other studies of school effectiveness to be related to school performance. These interpretations are further complicated when school comparisons are being made due to inevitable differences between schools on these factors. An accounting of these factors and their influence upon the comparisons made is missing from the reports appearing in the *Journal* (Note 1). In their absence, one principal commented:

"My biggest concern is that people believe the scores mean more than they do. To accurately compare schools, you need a huge amount of information because there are so many parameters, and I do not think we are close to having all that". (C. Lund, personal communication, May 9, 1990).

Third, the report is remiss in not reporting or otherwise ignoring the variability inherent in the data. If taken into account, as shown in Figure 1 for English 30 the conversion of marks to the school ranks reported in the *Journal* is not supportable. As suggested earlier on page 2, proper account must be made of the fallibility inherent in test scores.

District personnel, educational leaders, the media, and the public and parents need complete, valid, and accurate information if discussions about education and schooling are to produce positive changes in schools. Information that is incomplete in form and content will be counterproductive to any educational effort.

STANDARD 3.11 - When test-taking strategies that are unrelated to the constructs or content measured have been found to influence test performance significantly, these strategies should be explained to test takers before the test is administered either in an information booklet or, if the explanations can be made briefly, along with the test directions. The use of such strategies by all test takers should be encouraged if their effect facilitates performance and discouraged if their effect interferes with performance. (*Primary*). (American Psychological Association, 1985, pp. 27-28)

Illustration 2.

Test-wiseness is a cognitive ability or set of skills which a test taker can use to improve a test score no matter what the content area of a test (Sarnacki, 1979; Benson, 1988). If a test taker possesses test-wiseness, and if the examination contains susceptible items, then the combination of these two factors can result in an improved score; in contrast, a student low in test-wiseness will tend to be penalized every time he or she takes a test which includes test-wise components. Thus, a potential validity problem exists when one attempts to interpret the meaning of the test score. If test scores can be influenced by students being more or less test-wise, then those individuals involved with test development, administration, and interpretation need to carefully consider the construct of test-wiseness and how it affects scores.

A number of authors have offered definitions of test-wiseness (Diamond & Evans, 1972; Ebel & Damrin, 1960; Gibb, 1964; Stanley, 1971; Thorndike, 1951; Vernon, 1962), but for the purposes of this study, the definition proposed by Millman, Bishop and Ebel (1965) has been adopted: "a subject's capacity to utilize characteristics and formats of the test and/or test taking situation to receive a high score" (p. 707). Millman et al. further noted that test-wiseness is logically independent of an examinee's knowledge of the subject matter measured by the test's items. Basically then, test-wiseness encompasses both the method of measurement (flawed test items which provide test-wise cues) and characteristics of the test taker (a cognitive ability or set of abilities that an examined might employ in any testing situation regardless of the content measured).

Millman et al. (1965) included with their definition a taxonomy of test-wiseness principles which has served as the general framework for further studies of test-wiseness. Briefly, this taxonomy is organized into two categories. Part I contains elements applicable in most testing situations and which are independent of the test maker or test purpose. If employed, these strategies will help examinees avoid losing points for reasons other than lack of knowledge of the content tested. The principles listed in Part II of the taxonomy may prove

beneficial when the test taker has knowledge of particular test making behaviors or knowledge of particular testing practices gained from past experiences with tests similar in purpose and format.

Although it might be expected that standardized tests developed by professionals would be relatively immune to test-wiseness, research conducted in the United States suggests that this is not the case (Bangert-Drowns, Kulik, & Kulik, 1983; Benson, 1988; Sarnacki, 1979; Slack & Porter, 1980; Smith, 1982). These findings make it even more important to examine the influence of test-wiseness on provincial government examinations, particularly in view of the serious decisions that are made based on scores from these examinations.

Table 2

Summary of Test-Wiseness Analysis

Provincial Examinations		Test-wise Susceptable Items				Nonsusceptable Items			
	n	*k*	*Mean*	*SD*	*rel*[a]	*k*	*Mean*	*SD*	*rel*[a]
English 12	736	21	67.9%	13.9%	.58	7	58.8%	19.7%	.20
Algebra 12	267	25	62.5%	15.6%	.71	33	49.1%	14.7%	.73
Geography 12	166	43	59.5%	11.8%	.67	16	43.1%	14.0%	.48
History 12	208	39	71.7%	16.7%	.85	10	58.2%	20.1%	.46
Biology 12	261	37	64.6%	15.6%	.80	15	50.8%	18.1%	.57
Chemistry 12	174	28	69.6%	15.4%	.76	20	59.8%	17.6%	.69

Note. The difference between the mean performance on susceptible test-wise items and the mean performance on the nonsusceptible items is significant at the .01 level for each examination (correlated *t* test $df = n\text{-}1$).

[a] Internal consistency (Hoyt, 1941).

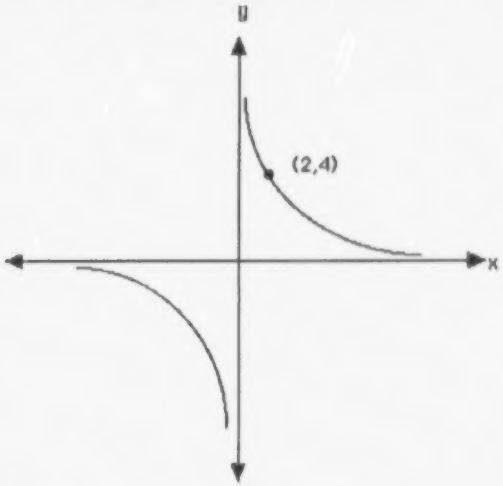
Shown in Table 2 are preliminary results of a study of the influence of test-wiseness upon student performance on the Grade 12 provincial examinations administered in British Columbia (Rogers & Bateson, 1991). Clearly the results speak for themselves. Provided in Table 3 are four examples of flawed test items from the Algebra 12 examination which contain test-wise cues.

Table 3

Sample Items Susceptible to Test-Wiseness

Algebra 12 N = 267

Use the following diagram to answer question 7.



7. Which of the following could be the equation of the function whose graph is shown above?

		p	r_{pbis}	$\overline{X}$	
A.	$x + 6 = 6$	3.7	-.19	24.20	Absurd option
B.	$y = x^2$	16.1	-.30	26.35	
C.	$y^2 = 8x$	10.1	-.21	26.70	
D.	$xy = 8$	70.0	.45	34.27	

Table 3 (Continued)

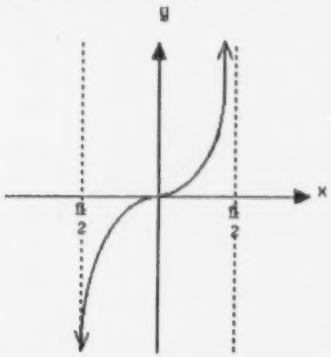
22. Which of the following is an arithmetic sequence?

A.	6,12,24	5.6	-.25	23.20	Similar option
B.	6,3,0	90.3	.26	32.54	
C.	6,8,12	3.0	-.11	26.63	Similar option
D.	6,1,-5	1.1	.01	32.67	

24. Determine the value of the common ratio for the sequence -8, 4, -2....

	p	r_{pbis}	$\overline{X}$		
A.	-4	0.0	0.00	0.00	
B.	-2	12.7	-.03	25.94	Opposite option
C.	-1/2	86.9	.29	32.76	Opposite option; stem-option
D.	12	0.4	-.07	22.00	

Use the following graph to answer question 37



37. Which of the following functions is partially illustrated by the above graph?

A.	y = cot x	11.6	-.16	28.29	Opposite option
B.	y = csx x	5.2	-.16	26.36	
C.	y = sec x	3.7	-.14	26.60	
D.	y = tan x	79.4	.28	33.01	Opposite option

115

The Need for Measurement Standards in Canada

The two illustrations briefly described in this paper raise serious concerns about the practice of measurement in Canada. The situation becomes even more acute when one considers the increase in testing taking place in Canada today. It is likely that the number of provinces with provincial examinations at various age or grade levels will increase or be expanded. The Council of Ministers of Education is at this moment assessing the proposals it received in response to its request for proposals for the development and validation of pan- Canadian tests of literacy and numeracy for 13- and 16- year-old students. There remains the possible introduction of a national scholarship examination for students entering first year science, engineering, or technology. Several Canadian provinces will be participating in the second International Assessment of Educational Progress; earlier this month provinces were asked to consider participation in Third International Study of Mathematics of the International Association for the Evaluation of Educational

What these results reveal is that the scores achieved on the provincial examination can not be simply and validily interpreted as a measure of the construct of interest. Instead, the scores of many of the students are confounded by the presence of test-wiseness.

Achievement (IEA). This listing does not touch upon testing conducted by school counsellors, school psychologists, and psychologists, or by school districts, school administrators, and classroom teachers (Wilson, 1990). In light of situations like those described above, can we be sure that the testing practices of these organizations and people yield reliable (dependable) and valid scores and score interpretations? I think not.

A Strategy for Improving Testing Practice in Canada

The call of this paper is that a committee comprised of representatives of participants in the testing process be formed in Canada to establish standards — principles generally accepted by professional associations as indicative of sound practice — for teaching and evaluation practice in Canada. Such a committee should work on behalf of persons and groups involved in delivering educational services. Its purposes should include promoting teaching and learning by identifying, articulating, and promoting the assimilation of measurement and testing standards which allow those involved in commissioning, conducting, using, and disseminating educational evaluations to use their judgment and creativity responsibly, but within the boundaries of "acceptable and sound measurement and evaluation practice."

Membership. The organizations represented on the committee called for should include associations whose members commission and use education testing and evaluations, associations whose members develop and disseminate tests and testing programs, and associations whose members are affected by or whose work is the subject of, testing and evaluations. Due regard must be given to both English and French representation. In Canada, these associations might include the Council of Ministers of Education, Canadian Educational Associa-

tion, Canadian School Trustees' Association, Canadian Society for the Study of Education, Canadian Guidance and Counselling Association, Canadian Psychological Association, Canadian Council for Special Education, and the Canadian Teachers' Federation, for example (Note 2).

Conclusion

The call of this paper is for a committee representative of the participants in the testing process in Canada to establish standards for the testing and evaluation practice in Canada. Educational and psychological testing can and has played a significant and important role in improving teaching and learning. The proper use of well-constructed and validated testing practices provides a better basis for making important decisions about individuals and programs than would otherwise be available. In consideration of recent criticisms (McLean, 1990, Rogers, 1990), the existence of faculty testing practices such as those illustrated earlier, and the increase in testing presently occurring in Canada, the intent of such a set of standards should be to provide a basis for evaluating the quality of testing practices as they affect the various parties involved.

Does the group of individuals here support this call?

Reference Notes

1. Although adjusting obtained test scores for the influence of demographic and opportunity-to-learn factors and then comparing schools is a frequently used technique for identifying effective and less effective schools, such a procedure may mask desired learning outcomes. For example, students from less affluent families should achieve at the same high level on the achievement tests as students from more affluent families (Brookover, 1987: Guskey & Kifer, 1990).

2. *The Standards of Educational and Psychological Testing* were produced by a committee appointed by the three sponsoring agencies. The American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

The Code of Fair Testing Practices in Education was developed by the Joint Committee on Testing Practices sponsored by the American Educational Research Association, American Psychological Association, National Council on Measurement in Education, American Association for Counselling and Development/Association for Measurement and Evaluation in Counselling and Development, and the American Speech-Language-Hearing Association.

There is in the United States a Joint Committee on Standards for Educational Evaluation. Established in 1975, this Joint Committee is currently sponsored by fifteen professional educational associations including the Canadian Society for the Study of Education with the assistance of the Faculty of Education at the University of Alberta, to promote concern for evaluation of high quality. The

work of this committee resulted in the publication of *The Standards for Evaluations of Educational Programs, Projects and Materials* and *The Personnel Evaluation Standards*. The Committee is now beginning to develop standards for the evaluation of students.

References

Alberta Education. (1989a). Provincial report, June 1989 administration: Diploma examinations programs. Edmonton, Alberta: Alberta Education.

Alberta Education. (1989b). Provincial Report, June 1989 administration: Achievement testing program. Edmonton, Alberta: Alberta Education.

American Psychological Association. (1988). *Code of fair testing practices in education*. Washington, DC: American Psychological Association.

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Bangert-Drowns, R.L., Kulik, J.A., & Kulik, C-L.C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research*, 53, 571-585.

Benson, J. (1988). The psychometric and cognitive aspects of test-wiseness: A review of the literature. *Phi Delta Kappan*.

Brookover, W.B. (1987). Distortion and overgeneralization are no substitutes for sound research. *Phi Delta Kappan*, 69, 225-227.

Diamond, J.J., & Evans, W.J. (1972). An investigation of the cognitive correlates of test-wiseness. *Journal of Educational Measurement*, 9, 145-1.

Ebel, R.L., & Damin, P.E. (1960). Test and examinations. In C.W. Harris (Ed.), *Encyclopedia of Educational Research* (3rd edition). New York: The MacMillan Company.

Gibb, B.G. (1964). Test-wiseness as a secondary cue response. (Doctoral dissertation, Stanford University). Ann Arbor, Michigan's University Microfilms, No. 64-7643.

Guskey, T.R. & Kifer, E.W. (1990). Ranking school districts on the basis of statewide test results: Is it meaningful or misleading? *Educational Measurement: Issues and practice*, 9, 11-16.

McLean, L.D. (1990). Time to replace the classroom test with authentic measurement. *Alberta Journal of Educational Research*. 36, 78-84.

Millman, J., Bishop, H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, 25, 707-726.

Panzeri, A. (April 17, 1990). Old scona is heads above public school counterparts. *The Edmonton Journal*.

Panzeri, A. (April 18, 1990). Test scores misleading, principals say. *The Edmonton Journal*, p.

Panzeri, A. (April 23, 1990). Teachers, principals criticize decision to make scores public. *The Edmonton Journal*, pp. B1-B2Province of Alberta. (1988). *School Act*. Edmonton, Alberta: Queens Printer for Alberta.

Rogers, W.T. (1990). Current educational climate in relation to testing. *Alberta Journal of Educational Research*, 36, 52-64.

Rogers, W.T. & Bateson, D.J. (1991). The influence of test-wiseness on performance of high school seniors on school leaving examinations. *Applied Measurement in Education*, 4 (2), 159-183.

Sarnacki, R.E. (1979). An examination of test-wiseness in the cognitive test. domain. *Review of Educational Research*, 49, 252-272.

Slack, W.V., & Porter, D. (1980). The Scholastic Aptitude Test: A critical appraisal. *Harvard Educational Review*, 50, 154-175.

Smith, J.K. (1982). Converging on the correct answer: A peculiarity of multiple-choice items. *Journal of Educational Measurement*, 19, 211-219.

Stanley, J.C. (1970). Reliability. In R.L. Thorndike (Ed.), *Educational Measurement* (Second edition). Washington, DC: American Council on Education.

Thorndike, E.L. (1951). Reliability. In E.F. Lindguist (Ed.), *Educational measurement*. Washington, DC: American Council on Education.

Vernon, P. (1962). The determinants of reading comprehension. *Educational and Psychological Measurement*, 22, 269-286.

Wilson, R.J. (1990). Processes in evaluating student achievement. *Alberta Journal of Educational Research*, 36, 4-17.